



Faculty of Computer Science and Information Technology
University of Malaya, Kuala Lumpur, Malaysia
Session 2002/2003

Perpustakaan SKTM

Data Mining in Electronic Commerce using Classification Technique

Prepared by:
Lim Lee Chin
WET 000011

Under the supervision of
Mr. Teh Ying Wah

And moderation of
Mr. Woo Chaw Seng

*A dissertation presented to the Faculty of Computer Science and Technology of
University Malaya in partial fulfillment of the requirements for the
Degree of*

Bachelor of Information Technology

Abstract

This thesis project is carried out in order to fulfill one of the academic requirements for Bachelor of Computer Science from the Faculty of Computer Science and Information Technology, University of Malaya. The project involved the development of *Data Mining in Electronic Commerce using Classification Technique*.

Firstly, this report will cover the general concept of data mining, project motivation, objectives of data mining, scope of project, current limitation, and an expected outcome and project schedule. Secondly, this report will review the literature. This part will cover basic knowledge of data mining system. Then followed by introducing data mining techniques and tools. Methodology part also will be discussed in chapter 3. In the system analysis, it is about analysis of functional and non – functional requirements specification. Next part is System Design that describes the data flow of my system. The final part of my project is system implementation and testing. It is description of the environment in which the systems are developed and implemented after completion. It also includes verifications and validations of the system to make sure the errors are in the minimum level.

Acknowledgement

First of all I would like to take this opportunity to extend my utmost gratitude to my supervisor for this thesis project, Mr. Teh Ying Wah for his guidance and supervision. The sound advices given throughout the project has helped tremendously.

Next, I would also like to thank Mr. Woo Chaw Seng, my moderator for this thesis project, for sharing his experience and provide constructive ideas, comments and thoughts during VIVA of this project.

Last but not least, I would also like to thank each and everyone that directly or indirectly contributed to the success of making this thesis report as well as the development of *Data Mining in Electronic Commerce*. Thank you all once again for your kind support, cooperation, invaluable guidance, precious advice, and many more.

Table of Contents

Abstract.....	i
Acknowledgement.....	ii
Table of Contents.....	iii
List of Tables.....	viii
List of Figures.....	ix

Chapter 1: Introduction

1.1 General Concept.....	1
1.2 Project Motivation.....	2
1.3 Objectives of Data Mining.....	3
1.4 Scope of Project.....	4
1.5 Current Limitation.....	5
1.6 Expected Outcome.....	6
1.7 Project Schedule.....	7
1.8 Summary of Chapter.....	8

Chapter 2: Literature Review

2.1 What is Data Mining?.....	10
2.2 Data Mining versus Traditional Database Queries.....	12
2.3 Data Mining Technology.....	13
2.3.1 Data Mining Algorithms.....	14
2.3.1.1 Classification.....	14
2.3.1.2 Association.....	17
2.3.1.3 Sequential Patterns.....	18
2.3.1.4 Clustering.....	19
2.3.2 Classification Technique.....	20
2.3.2.1 Naïve Bayesian Classifier.....	21
2.3.2.2 Decision Trees.....	23
2.4 Data Mining Applications.....	24
2.5 Future Challenges of Data Mining.....	26
2.6 What is Electronic Commerce?.....	28

2.6.1 Basic Function of Electronic Commerce Systems.....	29
2.7 E – Commerce and Data Mining.....	32
2.7.1 What E – Commerce Problems Does Data Mining Solve?.....	33
2.8 Data Mining Existing System.....	34
2.8.1 PolyAnalyst 4.4.....	34
2.8.1.1 New Features of PolyAnalyst 4.4.....	34
2.8.1.2 Machine Learning Algorithms.....	35
2.8.2 DBMiner.....	37
2.8.2.1 The Features of DBMiner.....	38
2.8.2.2 Current Products of DBMiner – DBMiner 2.0.....	39
2.8.3 Clementine – SPSS Inc.....	41
2.8.3.1 The Features of SPSS.....	42
2.8.4 IBM DB2 Intelligent Miner for Data.....	44
2.8.4.1 IBM DB2 Intelligent Miner for Data.....	45
2.8.4.2 Features and Benefits of DB2 Intelligent Miner for Data.....	45
2.8.5 Oracle9i Data Mining.....	46
2.8.5.1 Competitive Advantages of Oracle9i Data Mining.....	47
2.8.5.2 Supported Algorithms in Oracle9i Data Mining.....	47

Chapter 3: Methodology

3.1 The Waterfall Model.....	49
3.2 Prototyping Model.....	51
3.3 Why Choose Waterfall Model with Prototyping?.....	52
3.4 Data Gathering.....	55

Chapter 4: System Analysis

4.1 Requirement Analysis and Specification.....	57
4.1.1 Functional Requirements Analysis.....	57
4.1.2 Non – Functional Requirements Analysis.....	59
4.2 System Development Tools Analysis.....	61
4.2.1 Operating System.....	61
4.2.1.1 Windows 2000 Professional.....	61
4.2.1.2 Windows NT.....	62

4.2.2 Database Management.....	63
4.2.2.1 Microsoft Access 2000.....	63
4.2.2.2 Microsoft SQL Server 2000.....	64
4.2.3 Application Programming Language.....	64
4.2.3.1 JAVA.....	65
4.2.3.2 Microsoft Visual Basic 6.0.....	66
4.2.4 Application Programming Software.....	66
4.2.4.1 JBuilder 7.0.....	66
4.2.4.2 JCreator 2.5.....	67
4.2.4.3 TextPad.....	68
4.3 The Tools of Choices.....	69
4.3.1 Operating System – Windows 2000 Professional.....	69
4.3.2 Database Management – Microsoft Access 2000.....	70
4.3.3 Application Programming Language – JAVA.....	71
4.3.4 Application Programming Software – TextPad.....	72
4.4 System Requirement.....	73
4.4.1 Hardware Requirement.....	73
4.4.2 Software Requirement.....	73

Chapter 5: System Design

5.1 Architectural Design. of the Project.....	74
5.2 System Structure Chart.....	75
5.3 Data Flow Diagrams (DFD).....	76
5.3.1 Context Diagram.....	77
5.3.2 Diagram 0 (Level 0).....	78
5.4 Entity Relationship (ER) Diagram.....	79
5.5 Database Design.....	81
5.6 User Interface Design.....	82

Chapter 6: System Implementation

6.1 Development Environment.....85

6.1.1 Hardware Configuration.....85

6.1.2 Software Tools.....86

6.1.2.1 Software Tools for Development.....86

6.1.2.2 Software Tools for Report Writing.....86

6.2 Program Development.....86

6.2.1 Review the Program Documentation.....87

6.2.2 Design the Program.....87

6.2.3 Code the Program.....88

6.2.4 Test the Program.....88

6.2.5 Document the Program.....88

6.3 Program Coding.....89

6.3.1 Methodology.....89

6.3.2 Coding Principles.....89

6.3.3 Database Connectivity.....90

6.4 Database Implementation.....92

Chapter 7: System Testing

7.1 Unit Testing.....95

7.2 Integration Testing.....97

7.3 System Testing.....98

Chapter 8: System Evaluation & Conclusion

8.1 Problem and Solutions.....99

8.1.1 Problem and Solution During System Studies and Analysis.....99

8.1.1.1 Wide Area of Studies.....99

8.1.1.2 Determining the Project Scope.....99

8.1.2 Problems and Solution During System Implementation and Testing.....99

8.2 System Strengths.....100

8.3 System Limitations.....101

8.4 Future Enhancement.....101

8.5 Conclusion.....102

Bibliography 104

Installation Guide

ODBC Connection Guide

User Manual

Appendix A – Example Source Code

Softcopy

University of Malaya

List of Tables

<i>Table 1.1: Project Schedule</i>	7
<i>Table 4.1: Hardware Requirement</i>	73
<i>Table 4.2: Software Requirement</i>	73
<i>Table 5.1: Data flow diagram symbols</i>	76
<i>Table 5.2: ERD symbols</i>	79
<i>Table 5.3: ItemDetails Table</i>	81
<i>Table 5.4: CustomerDetails Table</i>	82
<i>Table 6.1: Hardware specification of computer used for system development</i>	86
<i>Table 6.2: Software tools for Development</i>	86

List of Figures

Figure 1.1: Project Schedule	7
Figure 2.1: KDD Process.....	10
Figure 2.2: Example of Classification.....	16
Figure 2.3: Example of Naïve Bayesian Classifier.....	23
Figure 2.4: Decision Tree Structure.....	24
Figure 2.5: Electronic – Commerce Model.....	31
Figure 2.6: Classify algorithm.....	36
Figure 2.7: Cluster algorithm.....	36
Figure 2.8: Decision Tree Exploration Engine.....	37
Figure 2.9: Decision Tree Classification.....	41
Figure 2.10: Clustering.....	41
Figure 2.11: Visual Association.....	41
Figure 2.12: OLAP (Summarizes).....	41
Figure 2.13: SPSS Data Editor.....	43
Figure 2.14: SPSS OLAP Cubes.....	44
Figure 2.15: SPSS Statistic Coach.....	44
Figure 2.16: Examination of Predictions and Classification.....	48
Figure 2.17: Results of Association Analysis.....	48
Figure 3.1: The Waterfall Model.....	50
Figure 3.2: The Prototype Model.....	52
Figure 3.3: Waterfall Model with Prototyping.....	53
Figure 5.1: System Structure Chart.....	75
Figure 5.2: Context Diagram for Data Mining in E – Commerce using Classification Technique.....	77
Figure 5.3: Diagram 0 for Data Mining in E – Commerce.....	78
Figure 5.4: ERD for Data Mining in E – Commerce using Classification Technique.....	79
Figure 5.5: Item Entity and its attributes.....	80

Figure 5.6: Customer Entity and its Attributes.....80

Figure 5.7: Login Module Interface.....83

Figure 5.8: Add New User Interface.....83

Figure 5.9: Main Module Interface.....84

Figure 5.10: Analyse Data Interface.....84

Figure 6.1: The five steps of Program Development.....87

Figure 7.1: Testing Stages.....94

Figure 7.2: Unit Testing.....96

Introduction

- 1.1 *General Concept*
- 1.2 *Project Motivation*
- 1.3 *Objectives of Data Mining*
- 1.4 *Scope of Project*
- 1.5 *Current Limitation*
- 1.6 *Expected Outcome*
- 1.7 *Project Schedule*
- 1.8 *Summary of Chapter*

Chapter 1: Introduction

1.1 General Concept

By referring to the books of Data Mining Techniques for Marketing, Sales and Customer Support by John Wiley & Son, the definition of Data Mining is stated as:

“Data Mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.”

In other word, data mining is the extraction or “mining” knowledge information from large databases, that can helps miner focus on the most important information in their data warehouses. Data mining tools allowing business to make proactive, knowledge driven decisions by predict future trends and behaviors.

The term ‘Data Mining’ is actually describing the concept of knowledge discovery in databases (KDD), which is a step in KDD process. KDD is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Data mining is the high – level application techniques used to present and analyze data for making decisions.

There are several data mining algorithms are used to solve particular problems that categorized as association, classifications, sequential patterns and clustering. Association analysis is the discovery of association rules showing attribute – value conditions that

occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. In the sequential patterns, the input data is set of sequences, called data – sequences. Each data sequences are an ordered list of transactions (or item sets), where each transaction is a set of items (literals). Typically there is a transaction – time associated with each transaction. A sequential pattern also consists of a list of sets of item. However, clustering analyzes data objects without consulting a known class label. It can be used to generate such labels. In the chapter two, I will review all these algorithms especially classification which is my task in developing this project.

1.2 Project Motivation

There are several factors to bring data mining to the forefront of electronic – commerce:

- ✓ The untapped value in large databases and lots of data is being collected and warehoused
- ✓ Consolidation of databases records towards a single customer view
- ✓ The ability to custom manufacture, advertise and market to individuals and small market segments
- ✓ Concept of an information or data warehouse, from the consolidation of databases
- ✓ The unknown true current condition of customer data

- ✓ The ways to compares customers' data to the average database in the industry
- ✓ Reduction in the data storage and processing cost that provide ability to collect and accumulate data
- ✓ Computing has become affordable/ powerful multiprocessor computers
- ✓ Competitive pressure is strong which provide better, customized services for an edge and information becoming product in its own right

1.3 Objectives of Data Mining

The objectives of developing data mining are outline as followings: -

- To extract useful information from the database to process the massive volume of data effectively for product planning purposes
- To identify the trends, deviations, relations of the products so that an effective decision can be made
- To finds patterns and relationships in data by using sophisticated techniques to build models. A good model is a useful guide to understanding business and making decision
- To discover meaningful patterns and rules from large quantities of information by automatic or semiautomatic means
- To increase profits and reduce go – to – market costs
- To explore high – dimensional databases to help clients better understand complex purchasing behavior

1.4 Scope of Project

The scope of this project can be divided into three scopes, which are Administrator's Scope, User's Scope, and System's Scope.

Administrator's Scope

- The database administrator should be able to manage the DBMS. The administrators can also update or retrieve the data that stored inside the database to maybe generate a performance report or just to check other information.
- The analyst should occasionally help in setting up discovery models.
- The analyst should be able to analyze the data collected into meaningful patterns and rules to make business decisions.
- The analyst should get more information after they have seen some results
- Administrators can register the customers by providing them with username and password in order for them to access the packages tracking system.

User's Scope

- The users should be empowered with direct, on – demand access to refined knowledge. They should be given consistent and correct answers without knowing the statistics. Besides that, the system interface should be as easy – to – use as a Web – browser.
- The users have the rights to access to the system to see about the benefit, quality and usability of the system.
- The users should get refined information when they need it.

System's Scope

- The system should work on databases with a large number of records. Therefore, the best option is for a data mining system to work on the real databases and not on samples, extracts or flat files.
- The system should be able to handle moderate to large volumes of data on a powerful server.
- The system must cover a wide range of patterns and provide high quality of information.

1.5 Current Limitation

There are a few of limitations of data mining technology:-

❖ *Cost, Time and Effort*

The data mining setup can be expensive running. Many man – hours of development are needed, involving complicated procedural steps and product choices. There is a need for data scrubbing or cleaning programs, and there is no single high – powered system that can handle this. Some of the data mining functions involve steep learning curves for the end – users, since higher computing power is directly related to the depth of knowledge on how the data mining system actually works. Writing SQL queries can be complex and difficult, even with a Windows – based front end tool. Extensive training and practice are still needed for most users.

❖ *Low - end software*

Some of the lower – end software available for data warehouse analysis tools are available for thousands of dollars, but these are piece – meal modules. These have limited query capabilities and its inability to perform multidimensional analyses – impossible to ask open – ended questions to find associations between data items. Making additions, changes and other replacements to these smaller software systems can lead to integration and implementation problems. Many of the current data mining methods are not interactive and cannot incorporate prior knowledge about a problem except in simple ways.

❖ *Large databases*

The large size of business databases presents problems in terms of finding efficient algorithms for association rules. Large numbers of fields (or attributes) also increases the need for search space enormously, and in addition, it increases the chances that the dynamic nature of its data – variables maybe modified, deleted or augmented with new measurements over time.

1.6 Expected Outcome

The development of Data Mining in Electronic Commerce will be expecting some of the outcomes listed as following: -

- ✓ Able to build an user – friendly and ease – to – use system that can help the data miner to analyze the data.

- ✓ Able to provide a user access pattern of Web documents stored in an information provider's web server.
- ✓ Able to discover meaningful patterns and rules from large quantities of information.
- ✓ Can leverage hidden information in its data
- ✓ Able to analyze and validate the results

1.7 Project Schedule

In order to reduce inherent uncertainty in determining the time estimations, the expected time of all the activities will be estimated optimistically. For the project schedule, please refer to *Table 1.1* and *Figure 1.1* below:

No	Task	Start Date	End Date	Duration
1	Research on Thesis Title	10/6/02	5/7/02	3 weeks
2	Requirement Analysis	15/6/02	15/7/02	4 weeks
3	Literature Review	15/6/02	5/8/02	7 weeks
4	System Analysis and Design	9/7/02	15/9/02	10 weeks
5	System Development	1/10/02	6/1/03	13 weeks
6	Implementation and Testing	3/12/02	20/1/03	5 weeks
7	Documentation	18/6/02	27/1/03	31 weeks

Table 1.1: Project Schedule

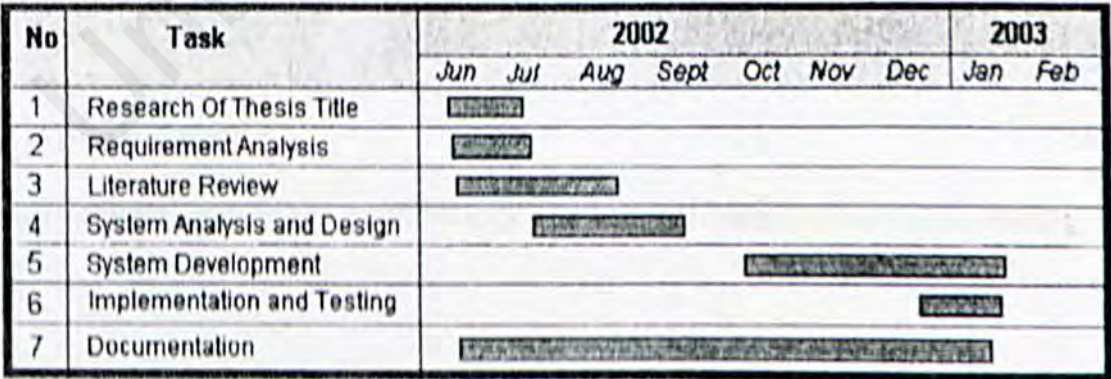


Figure 1.1: Project Schedule

1.8 Summary of Chapter

The purpose of this report is to document all the essential and pertinent information collected and applied in developing this project. It covers many phases that includes several instructional design phases that included analysis, design and development phase of the system. Below is the report summary: -

Chapter 1: Introduction

This chapter gives an introduction to the entire project. This chapter outlines the brief history of data mining, objectives, scopes of project and limitations and expected outcome. It also features the project schedule for this project.

Chapter 2: Literature Review

In this chapter gives a detail explanation on topics researched and studied that are relevant to this project. This part will focus on data mining techniques and related data mining with electronic commerce. It also analyses the meaning of classification technique which I am in charge in this part after separated four techniques among four students.

Chapter 3: Methodology

This chapter emphasized on the methodology and how information and requirements is gathered.

Chapter 4: System Analysis

This chapter is about analysis of functional and non – functional requirements specification. Besides that, it also analysts the development tools that available and then choose the best tools or software to develop the system.

Chapter 5: System Design

This chapter will focus on the system design phases such as the system's architecture, the conceptual and technical design processors of the system. It will include database design and generated report.

Chapter 6: System Implementation

This chapter gives a description of the environment in which the systems are developed and implemented after completion.

Chapter 7: System Testing

This chapter gives a description whether the system functional as its requirements and specification. It includes verifications and validations of the system to make sure the errors are in the minimum level.

Chapter 8: System Evaluation and Conclusion

This chapter gives an evaluation about the system in terms of the problems and it's solutions, strengths and limitations of the system development, suggestions for further enhancements and conclusion.

Literature Review

- 2.1 *What is Data Mining?*
- 2.2 *Data Mining versus Traditional Database Queries*
- 2.3 *Data Mining Technology*
- 2.4 *Data Mining Applications*
- 2.5 *Future Challenges of Data Mining*
- 2.6 *What is Electronic – Commerce?*
- 2.7 *E – Commerce and Data Mining*
- 2.8 *Data Mining Existing System*

Chapter 2: Literature Review

The literature review is the most important part of **any project**. This is because it places the project in the context of other existing and **similar projects** or with similar characteristics. Literature review helps to generate an overall idea and gives an accurate and useful view of existing systems and the features that the existing systems have. It is important to review the existing systems so that the drawbacks and the plus points of the system can be identified and used to design a system which is more perfect. It also is a mean of delivering the knowledge of the strengths and limitations of development tools that best suits to develop proposed system.

2.1 What is Data Mining?

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules (John Wiley & Sons, 1997). Data mining is a step in KDD (Knowledge Discovery in databases) process which KDD is a non – trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. *Figure 2.1* below show the KDD process.

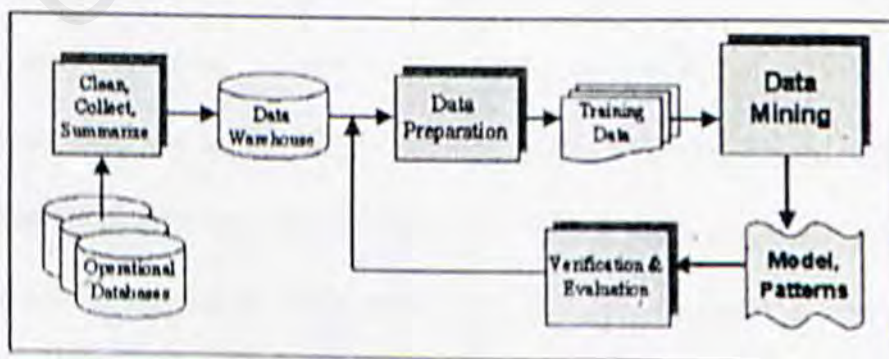


Figure2.1: KDD process

Data mining is a set of automated techniques used to extract buried or previously unknown pieces of information from large amounts of databases. Successful data mining makes it possible to unearth patterns and relationships, and then use this “new” information to make proactive knowledge – driven business decisions. Data mining then, centers on the automated discovery of new facts and relationships in data. The raw material is the business data, and the data mining algorithm is the excavator, sifting through the vast quantities of raw data looking for the valuable nuggets of business information.

Data mining is usually used for four main purposes: -

- (1) to improve customer acquisition and retention;
- (2) to reduce fraud;
- (3) to identify internal inefficiencies and then revamp operations
- (4) to map the unexplored terrain of the internet.

There are several popular types of tools used in data mining: neural network, decision trees, rule induction, and data visualization, which will briefly discuss later.

Data mining potential can be enhanced if the appropriate data have been collected and stored in a data warehouse – a system for storing and delivering massive quantities of data. Data warehousing is the process of extracting and transforming operational data into informational data and loading it into a central data store or warehouse. The promise of data warehousing is that data from disparate databases can be consolidated and managed from one single database.

There are three basic steps in data mining: -

Step 1: Data preparation, referred to as “scrubbing the data.” Data is selected, cleansed, and preprocessed under the guidance and knowledge of a domain expert.

Step 2: A data mining algorithm is used to process the prepared data, compressing and transforming it to make it easy to identify any latent valuable nuggets of information.

Step 3: The data mining output is evaluated to see if the mining algorithms discovered additional domain knowledge.

Data mining differs from other analytical tools in the approach used in exploring the data relationships. Traditional database queries can answer questions like “what were my car sales in Scotland in 1995?” Other analyses, often called multidimensional or online analytical processing, allow users to do more complex queries, such as comparing sales relative to plan by quarter and region for the prior two years. In both cases, however, the results are simply figures extracted from the data or an aggregate of existing data. The user, who, by framing the proper question, obtains the desired answer, already knows the relationship among these data.

2.2 Data Mining versus Traditional Database Queries

Traditional database queries contrasts with data mining since these are typified by the simple question such as “what were the sales of orange juice in January 1995 for the Boston area?”. Multidimensional analysis, often called on – line analytical processing (OLAP), lets users do much more complex queries, such as the comparison of actual and planned sales by region for the previous years of interest. Again, the emphasis in both

these cases is that the derived result(s) are value(s), which are an extraction or aggregation of existing data. Data mining, on the other hand, though the use of specific algorithms or search “engines”, attempts to source out discernable patterns and trends in the data and infers rules from these pattern. With these rules or functions, the user is then able to support, review and examine decisions in some related business or scientific area.

Companies are beginning to realize that their most valuable asset is information they possess on customer and buying patterns. Competitiveness increasingly depends on the quality of decision making from past transactions and decisions. The ability to improve the knowledge on customers and markets will enable businesses to better target their products and services. For example, retailers are able to target customers for sales promotions and manage disparate inventory databases; telecommunication companies can forecast demand patterns, profile and segment customer groups, customize billing and analyze profitability; and, financial institutions can consolidate information to segmentize financial products.

2.3 Data Mining Technology

The fundamental goals of data mining are prediction and description. Prediction makes use existing variables in the database in order to predict unknown or future values of interest. Description focuses on finding patterns describing the data and the subsequent presentation for user interpretation. The relative emphasis of both prediction and

description varies on the data mining system used. Several algorithms fulfill these objectives. These algorithms are incorporated into the various data mining methods.

2.3.1 Data Mining Algorithms

There are several data mining algorithms which are used to solve specific problems or objectives. There are categorized as classifications, associations, sequential patterns and clustering. The basic premise of associations is to find all generation develops profiles of different groups. A sequential pattern identifies sequential patterns subject to a user – specified minimum constraint. Clustering segments a database into subsets or clusters. However, the method that will be used in this project is classification technique.

2.3.1.1 Classification

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known) (Jiawei Han & Micheline Kamber, 2001). Example of classification as shown as figure 2.2 in the next page.

Classification is the most commonly technique used in data mining, it employs a set of preclassified examples to develop a model that can classify the population of large records. This tools will fraud detection and credit – risk applications.

For a given set of records with its corresponding attributes, a set of tags (representing classes of records) and an assignment of a tag for each record, a classification function examines the set of tagged records and produces descriptions of the characteristics of records for each classes. For example in credit analysis, the card issuing company will have customer records containing a number of descriptors. So for the customer with a known credit history, the customer's record is tagged as 'triple A or excellent'; 'good', 'medium' or 'poor'. The classification rule could be: -

- Customers with excellent credit history have a debt/equity ratio of less than 10%.

This rule could then be used to apply to new data sets for classification. Another example would be target marketing. Any company, which intends to carry out promotional mailings and using a profile generator, a classification or profile is developed characterizing the people who had responded to the previous mailing. This profile is then taken as a predictor of response to the current mailing. The mailing list is filtered such that the promotional materials are targeted towards those who match the profile. Besides target marketing and credit approval, profile generation is used for attached mailings and treatment – appropriateness determination.

There are four steps in the classification method: -

- i. Collection of the relevant set of data and partitioning of the data into training and testing data
- ii. Analysis of the relevance of the dimensions involved
- iii. Construction of the classification tree
- iv. Testing the effectiveness of the classification using the test data set

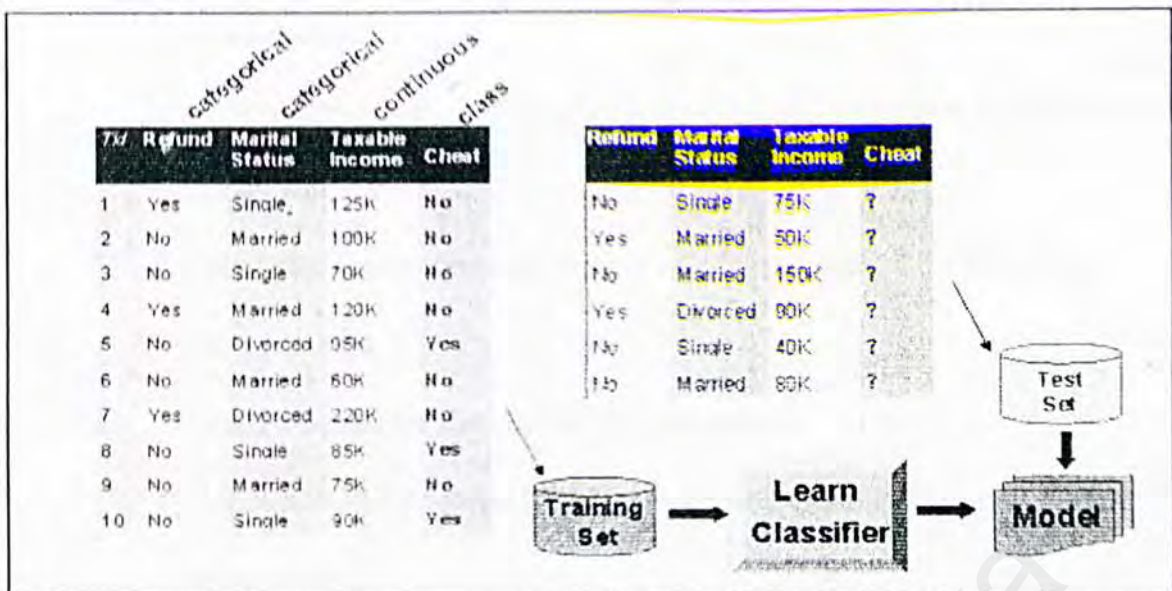


Figure 2.2: Example of Classification

Classification: Application

□ Direct Marketing

Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell – phone product.

Approach: 1) Use the data for a similar product introduced before

- 2) We know which customers decided to buy and which decided otherwise. This {buy, don't buy} decision forms the class attribute
- 3) Collect various demographic, lifestyle, and company – interaction related information about all such customers.
- 4) Use this information as input attributes to learn a classifier model

□ Fraud Detection

Goal: Predict fraudulent cases in credit card transactions.

Approach: 1) Use credit card transactions and the information on its account – holder

as attributes.

- when does a customer buy, what does he buy, how often he pays on time, etc.

2) Label past transactions as fraud or fair transactions. This forms the class attribute

3) Learn a model for the class of the transactions

4) Use this model to detect fraud by observing credit card transactions on an account

□ Customer Attrition / Churn

Goal: To predict whether a customer is likely to be lost to a competitor.

Approach: 1) Use detailed record of transactions with each of the past and present customers,

To find attributes

- how often customer calls, where he calls, what time – of the day he calls most, his financial status, marital status, etc.

2) Label the customers as loyal or disloyal

3) Find a model for loyalty

2.3.1.2 Association

This association algorithm has numerous applications, including supermarket, inventory planning, shelf planning, attached mailing in direct marketing and promotional sales planning. For example, the association rule derives, from data mining a database of transactions (via the product bar – code scanner), a 'market basket' or list consisting of

the set of items bought by a customer on a single visit to a store. The association rule could be: -

- 75% of customers who buy corn chips

The number '75%' refers to the confidence factor, a measure of the predictive power of the rule. The left hand side (LHS) item is Coke, whereas corn chips are the right hand side (RHS) item, of the rule. The algorithm procedures a large number of these rules and it is up to the user to select the subset of rules that have higher confidence levels and the percentage of the lists or 'market baskets' that allow this rule. There might also be multiple associations such as: -

- 65% of customer who buy Coke and corn chips also buy salsa.

It is important for the user to determine whether there is some element of chance correlation (Coke and chips were on sale) or whether there is an unknown but important correlation (salsa was also being bought). The impact here is how does the supermarket boost salsa sales? What happens if there is a Pepsi promotion? Conversely, what item should be stacked with each other on the same shelf; inventory of related items should closely follow each other.

2.3.1.3 Sequential Patterns

The input data is set of sequences, called data – sequences. Each data sequences are an ordered list of transactions (or item sets), where each transaction is a set of items (literals). Typically there is a transaction – time associated with each transaction. A sequential pattern also consists of a list of sets of items. The problem is to find all

sequential patterns with a user – specified minimum support, where the support of a sequential pattern is the percentage of data sequences that contain the pattern.

This technique looks at purchases, or events, occurring in a sequence over time. For example, retailer might discover that customers, who buy TVs, tend also to purchase 8mm camcorders 60% of the time.

- 60% of customers buy TVs followed by 8mm camcorders.

A similar sequence rule could be: -

- 90% of the time, whenever a sale of Coke goes up, sales of pretzels also goes up.

This will impact greatly on store layout and also identify customers who can be targets of sales promotion efforts for camcorders, if they have purchased TVs within the last 3 months. This type of algorithm is especially useful for catalog companies and financial investment firms, who are able to analyze sequences of events that affect the prices of financial instruments.

2.3.1.4 Clustering

Clustering is a process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self – similarity. It is up to the miner to determine what meaning,

if any, to attach to the resulting clusters. A particular cluster of symptoms might indicate membership in different subcultures.

Clustering will segment a database into subsets or clusters. These can be created statistically or by using neural and symbolic unsupervised induction methods. The various neural and symbolic methods are distinguished by the type of attribute values that can be accepted (numeric, nominal and structured objects); cluster representation, and cluster organization (in a hierarchy or flat list). This segmentation approach was developed to deal with the processing of consumer surveys. For example, a survey questionnaire containing 25 multiple choice questions, can be analyzed by question – i.e. 35% answered response 'B' for question 1, and so forth. The challenge is to analyze this questionnaire as a collection of 25 answer patterns, each provided by a single consumer. This technique will divide consumers according their answer patterns, thus creating a set of groups which have the maximum similarity within them and the maximum difference between them. Some of the pertinent uses have been in the analysis of patent databases by the degree that there is agreement in the use of key words; analyzing text for concepts; understanding types of consumes in consumer surveys and finding relevant research articles.

2.3.2 Classification Technique

There are several types or technique that can use to analyze the classification pattern such a Naïve Bayesian Classifier, Neural Network, Decision Tree, Decision Table and so on. But, in this report I will discuss on Naïve Bayesian Classifier and Decision Tree.

2.3.2.1 Naïve Bayesian Classifier

Bayesian classifier is statistical classifier and they can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. It is based on Bayesian theorem described below. Studies comparing classification algorithms have found simple Bayesian classifier known as Naïve Bayesian classifier to be comparable in performance with decision tree and neural network classifiers and they have also exhibited high accuracy and speed when applied to large databases.

Naïve Bayes classifiers assume that the effect of an attribute value on a given class is independent of values of other attributes. This assumption is called class conditional independence. It is made to simplify the computation and in this sense considered to be “Naïve”.

Bayes Theorem

Let X be the data sample whose class label is unknown. Let H be some hypothesis, such as data sample X belongs to a specified class C . For classification we want to determine $P(H/X)$, the probability that the hypothesis H holds given the observed data sample X .

$P(H/X)$ is the posterior probability. Or a posteriori probability, of H conditioned on X . For example, suppose world of data samples consists of fruits, described by their color and shape. Suppose that X is red and round, and that H is hypothesis that X is an apple. Then $P(H/X)$ reflects our confidence that X is an apple given that we have seen X is round and red. In contrast $P(H)$ is the prior probability, or a priori probability, of H . In

this example $P(H)$ is the probability that any given data sample is an apple, regardless of how the data sample looks. The posterior probability, $P(H/X)$, is based on more information (such as background knowledge) than prior probability, $P(H)$, which is independent of X .

Similarly, $P(X/H)$ is posterior probability of X conditioned on H . That is, it is the probability that X is red and round given that we know that it is true that X is an apple. $P(X)$ is prior probability of X i.e., it is the probability that a data sample from our set of fruits is red and round. Bayes theorem is useful in that it provides a way of calculating the posterior probability, $P(H/X)$, from $P(H)$, $P(X)$, and $P(X/H)$. Bayes theorem is

$$P(H|X) = P(X|H)P(H)/P(X)$$

Play Tennis - example

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

$P(p) = 9/14$
$P(n) = 5/14$

outlook	
$P(\text{sunny} p) = 2/9$	$P(\text{sunny} n) = 3/5$
$P(\text{overcast} p) = 4/9$	$P(\text{overcast} n) = 0$
$P(\text{rain} p) = 3/9$	$P(\text{rain} n) = 2/5$
temperature	
$P(\text{hot} p) = 2/9$	$P(\text{hot} n) = 2/5$
$P(\text{mild} p) = 4/9$	$P(\text{mild} n) = 2/5$
$P(\text{cool} p) = 3/9$	$P(\text{cool} n) = 1/5$
humidity	
$P(\text{high} p) = 3/9$	$P(\text{high} n) = 4/5$
$P(\text{normal} p) = 6/9$	$P(\text{normal} n) = 2/5$
windy	
$P(\text{true} p) = 3/9$	$P(\text{true} n) = 3/5$
$P(\text{false} p) = 6/9$	$P(\text{false} n) = 2/5$

Given a training set, we can compute the probabilities.

Outlook	P	N	Humidity	P	N
sunny	2/9	3/5	high	3/9	4/5
overcast	4/9	0	normal	6/9	1/5
rain	3/9	2/5			
Temperature			Windy		
hot	2/9	2/5	true	3/9	3/5
mild	4/9	2/5	false	6/9	2/5
cool	3/9	1/5			

An unseen sample $X = \langle \text{rain, hot, high, false} \rangle$

$$P(X|p) \cdot P(p) = P(\text{rain}|p) \cdot P(\text{hot}|p) \cdot P(\text{high}|p) \cdot P(\text{false}|p) \cdot P(p) \\ = 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$$

$$P(X|n) \cdot P(n) = P(\text{rain}|n) \cdot P(\text{hot}|n) \cdot P(\text{high}|n) \cdot P(\text{false}|n) \cdot P(n) \\ = 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286$$

Sample X is classified in class n (**don't play**)

Figure 2.3: Example of Naive Bayesian Classifier

2.3.2.2 Decision Trees

Decision tree is a wonderful tool for data mining. This methodology uses a hierarchy of *if – then* statements (if condition then outcome) to classify data. There are two main types of decisions trees: classification trees and regression trees, but these two trees have a same structure. The advantage in this application is that it is faster and more understandable than neural networks. The *if – then* statements could also be complex, especially if the condition list is long. For instance, consider the following example.

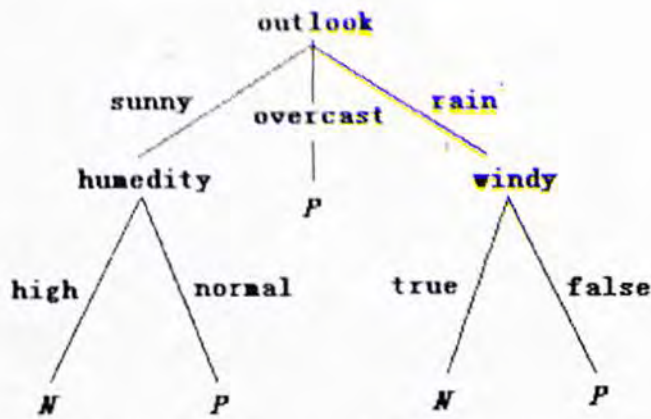


Figure 2.4: Decision Tree Structure

When to use Decision Trees

Decision trees methods are a good choice when the data mining task is classification of records or prediction of outcomes. Use decision trees when your goal is to assign each record to one of a few broad categories. Decision trees are also a natural choice when your goal is to generate rules that can be easily understood, explained, and translated into SQL or a natural language.

2.4 Data Mining Applications

Among the highest – profile users of data mining are the banking, financial, and telecommunications industries, but the full spectrum of users is very broad. A survey that have been done from Two Crows Corporation found these applications of data mining: -

- Ad revenue forecasting
- Churn (turnover) management
- Claims processing

- Credit risk analysis
- Cross – marketing
- Customer profiling
- Customer retention
- Electronic commerce
- Exception reports
- Food – service menu analysis
- Fraud detection
- Government policy setting
- Hiring profiles
- Market basket analysis
- Medical management
- Member enrollment
- New product development
- Process control
- Quality control
- Shelf management / store management
- Student recruiting and retention
- Targeted marketing
- Warranty analysis

2.5 Future Challenges of Data Mining

❖ *Data glut*

The current philosophy of data collection is 'collect now process later'. The issue arises as to whether we are collecting the right data in right amounts and how do we can distinguish between what is important / unimportant. It is thus essential to learn from data – in order to develop tools and techniques to determine the relevant data for collection, archival and purging.

❖ *Appropriate use*

There is a need for methodologies to assist in the strategic use of data for competitive advantage. For example, Canon found from data that Xerox is best in copier service – they changed rules of game by selling disposable copier cartridges – making copier service unnecessary. Some applications that predict market behavior or stock selection of mutual funds using neural networks may never be publicized if successful, since the goal is to attain competitive advantage.

❖ *Differing data types*

Data will not be confined to numeric data types only. Data processing will have to devise new algorithms to handle non – textual, geometric, multimedia data types – multi – lingual text, images, video, audio, graphical, temporal, relational, categorical and continuous data. There is a need for a unifying framework for

data representation and problem solving in order to learn and discover from large amounts of data.

❖ *Distributed environment*

As business evolve globally, data centers will be dispersed accordingly. The development of tools and techniques to enable data discovery from distributed databases and networked database environments, will be necessary. There should also be dynamic access to networked databases, perhaps even with the use of intelligent agents.

❖ *Direct interaction*

Data mining should be able to interact directly with the DBMS instead of allowing users to extract data from the DBMS and maintain these extracts outside the DBMS. This interaction will facilitate data mining operations and data management. If the data mining operation is done within the DBMS rather than in the application's memory space, the problem of scaling can be overcome.

❖ *Data structures*

Vendors are also debating whether it is better to set up a relational OLAP structure or a multidimensional one. Relational structures, where data is stored in tabular format, permits ad hoc queries and navigation by matching field values between tables. In a multidimensional architecture however, sets of cubes are arranged in arrays, with subsets created according to category. For example, in a sales database – these arrays can show data by geography, time and product.

2.6 What is Electronic Commerce?

Electronic commerce is an emerging model of new selling and merchandising tools in which buyers are able to participate in all phases of a purchase decision, while stepping through those processes electronically rather than in a physical store or by phone (with a physical catalog). The processes in electronic commerce include enabling a customer to access product information, select items to purchase, purchase items securely, and have the purchase settled financially.

The term 'electronic commerce' embraces electronic trading, electronic messaging, EDI, EFT, electronic mail (e-mail), facsimile, computer – to – fax (C – fax), electronic catalogues and bulletin board services (BBS), shared databases and directories, continuous acquisition and lifecycle support (CALS), electronic news and information services, electronic payroll, electronic forms (E – forms), online access to services such as the Internet, and any other form of electronic data transmission.

Internet commerce or E – commerce uses online electronic technology connected via the Internet to assist and enhance a variety of business processes, functions and systems. Using the Internet for commerce means far more than simply processing payment transactions when a customer buys products online, although this is how “ E – commerce” is now generally being defined.

Internet commerce systems are being established for automating and enhancing many aspects of communications, publishing, marketing, sales and customer service such as: -

- Customer Research
- Pre – Sales Enquiries
- Information Publishing and Dissemination
- Sales
- Advertising
- Promotions
- Public Relations
- Purchasing
- Transactions
- Funds Transfer
- Production
- Fulfillment
- Delivery
- After – Sales Service
- Ongoing Relationship Management and
- Customer Support

2.6.1 Basic Function of Electronic – Commerce Systems

Electronic commerce is coming of age. Electronic sales in a recent quarter are double those of the entire previous year. In some instances, companies create electronic – commerce capabilities out of a fear of falling behind competitors or as a result of the general momentum to expand the use of an existing Internet presence but the primary

value proposition is the prospect of increased revenue from new markets and creation of new, lower – cost, electronic – distribution channels.

Internet services providers (ISPs) are beginning to launch, or are at least evaluating, electronic – commerce hosting services. These services position the service provider as the customers' electronic – commerce capabilities, managing the networking and server aspects of the initiative. This allows the ISP's customers to concentrate on their core businesses and expands the relationship of the customer and the ISP. An ISP's ability to offer a rich electronic – commerce environment, on it own or in partnership with an electronic – business provider, will be important in differentiating high – value ISPs from lower – value, access – only ISPs.

Customer's Perspective

From a customer's perspective, the purpose of an electronic – commerce system is to enable customer to locate and purchase a desired good or services over the Internet when the customer is interested in making the purchase.

Merchant's Perspective

From a merchant's perspective, the key function of an electronic – commerce system is to generate higher revenues than the merchant would achieve without the system. In order to achieve this purpose, the electronic – commerce system must recreate or utilize existing data and business processes. All of the same processes that the merchant must have in place to support an in – store or catalog purchase must also be in place for an electronic purchase: product information, inventory systems, customer services, and

transaction capabilities (including credit authorization, tax computation, financial settlement, and shipping).

Additional functions of an electronic – commerce system, related to revenue generation, are to help redefine and enhance an enterprise’s brand strength, customer – service capabilities, and supply – chain effectiveness. An electronic – commerce system is one of the areas of an enterprise’s infrastructure that is open to customers via the Web, but it should be linked with other information technology (IT) systems that affect customer service (i.e., inventory and billing).

Basic Components

Provision of this basic system requires Internet access and an access device at the location of the home shopper, a Web – application server and electronic – commerce software (enabling catalog creation and transaction processing), security gateways to limit external access to internal data systems, and integration software to pull data from the appropriate support systems into the commerce environment (see Figure 2.5).

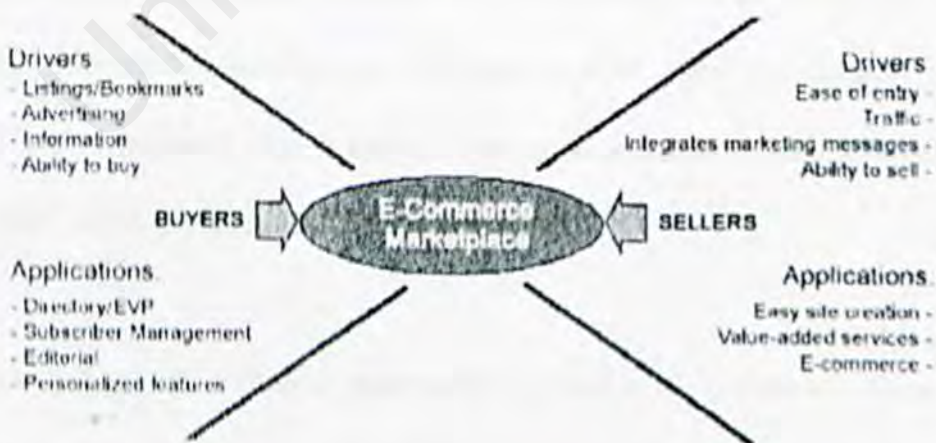


Figure 2.5: Electronic – Commerce Model

2.7 E-Commerce and Data Mining

Will electronic commerce be the *killer app* for data mining? E – commerce sites collect massive amounts of data on customer purchases, browsing patterns, usage times, and preferences. Information on competitors' offerings and prices also can be collected from the sites. They can adjust their prices and promotions quickly and systematic, based on changing trends and personalization rules. Because e – business implement close – loop computerized solutions, many of the traditional barriers to the effective application of data mining are significantly lower, such as access to data, data transformations, process automation, and timeliness of discoveries.

All areas of electronic commerce are relevant. Particular problems of interest include, but are not limited to: personalization (both model discovery and deployment), mass customization, increasing market basket value (e.g., cross – selling), improving customer satisfaction and loyalty, improving search facilities, recommender systems (e.g., collaborative filtering), improving navigation, improving marketing, improving advertising (e.g., ad matching and profiling), increasing frequency of visits and conversion rates, reducing costs, business – to – consumer and business – to – business transactions, competitive intelligence, shopping agents, and the transfer of mined knowledge to conventional stores and conventional distribution channels (e.g., direct channels, self – service channels, indirect channels).

Also relevant are general technical issues when applied to e – commerce. These include, but are not limited to: integration with large e – commerce systems and data warehouses,

incorporating performance feedback (e.g., campaign management) to improve models, data transformations (e.g., creation of customer signatures and profiles), multi – level data (e.g., hierarchical data), text mining, click stream mining (e.g., web log analysis and abstractions), integration with syndicated data, incorporating prior business knowledge, post – processing operations (e.g., visualization and workflow integration), privacy issues, and emerging standards (e.g., APIs).

2.7.1 What E– Commerce Problems Does Data Mining Solve?

Data mining can be used to solve almost any e – commerce problem that involves data, including: -

- Increasing business unit and overall profitability
- Understanding customer desires and needs
- Identifying profitable customers and acquiring new ones
- Retaining customers and increasing loyalty
- Increasing ROI and reducing costs on promotions
- Cross – selling and up – selling
- Detecting fraud, waste and abuse
- Determining credit risks
- Increasing Web site profitability
- Increasing store traffic and optimizing layouts for increased sales

2.8 Data Mining Existing System

A research was done to find various existing data mining system due to develop the Data Mining in Electronic Commerce. From the many types of online data mining system, the following existing systems would be discussed.

2.8.1 PolyAnalyst 4.4 (www.megaputer.com)

MEGAPUTER Intelligence Inc. announced the release of PolyAnalyst 4.4, a high – end Data Mining tool offering new features that dramatically simplify the process of integrating the results obtained through PolyAnalyst 4.4 data analysis in any external decision support applications and help processing very large databases in their entirety.

The new PolyAnalyst Model Application Wizard allows simple scoring of data in any external source through a standard SQL – based protocol, OLE DB. PolyAnalyst can export created models in XML/PMML format for store these models in a new OLE DB for Data Mining format suitable for direct data scoring. Additionally, PolyAnalyst 4.4 provides In – Place Data Mining capability for working with very large data in dynamic SQL mode through an intuitive interface.

2.8.1.1 New Features of PolyAnalyst 4.4

PolyAnalyst version 4.4 delivers numerous new features including: -

- Model Application Wizard for scoring external data in any source through a standard OLE DB for Data Mining protocol
- In – Place SQL – mode Data Mining for processing very large databases. Supported algorithms like Decision Tree, Market Basket Analyst and Clustering.

- Exporting created models to XML/PMML format
- Optimized Decision Tree algorithm based on Information Gain criteria
- Support for weighted accuracy classification into different classes in the Classify and Decision Tree algorithms
- Support for SAS format data files
- Improved Decision Tree viewer and exporting DT models in HTML format for comprehensive printing

2.8.1.2 Machine Learning Algorithms

PolyAnalyst provides the following machine learning algorithms for data exploration: -

- Classify
- Cluster
- Decision Forest
- Decision Tree
- Discriminate
- Find laws
- Link Analysis
- Market Basket Analysis
- Memory based Reasoning
- PolyNet Predictor
- Text Analysis
- Transactional Basket Analysis

PolyAnalyst 4.4 Screen Shots:

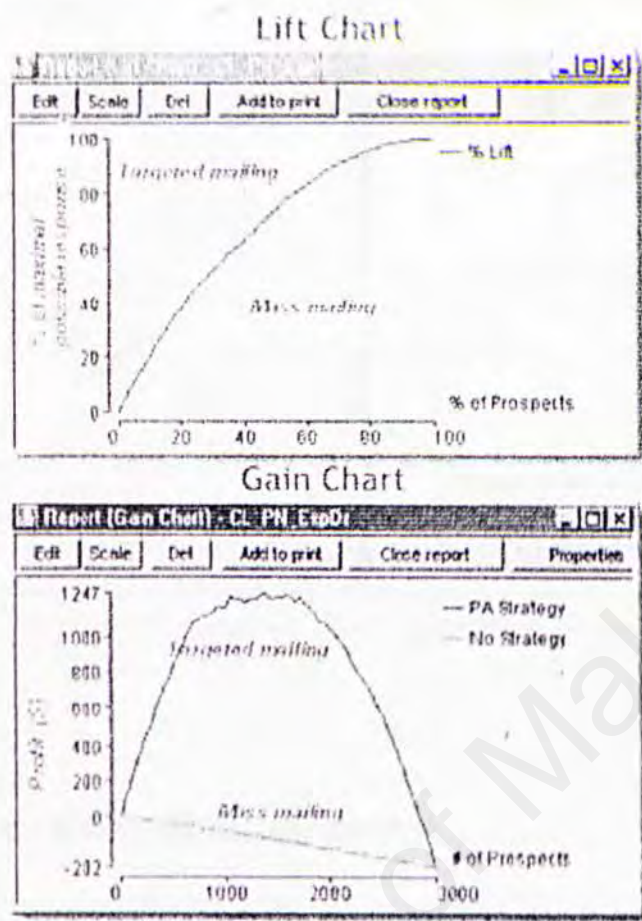
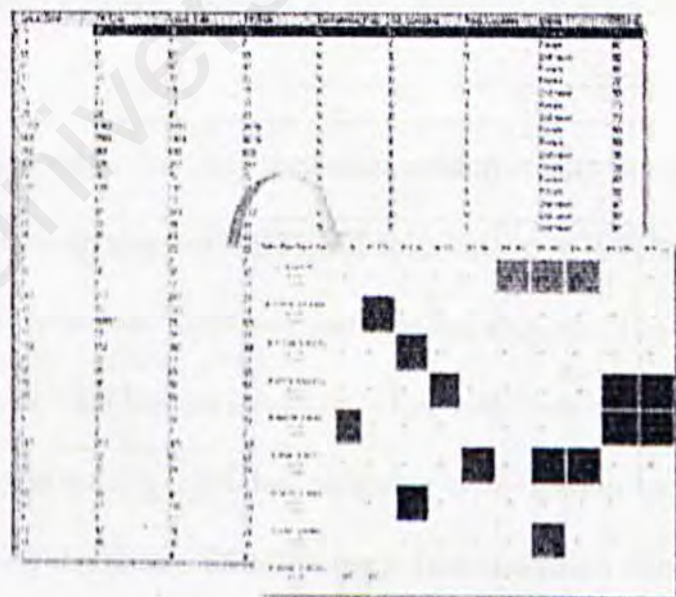


Figure 2.6: Classify algorithm



Groups of similar records
Figure 2.7: Cluster algorithm

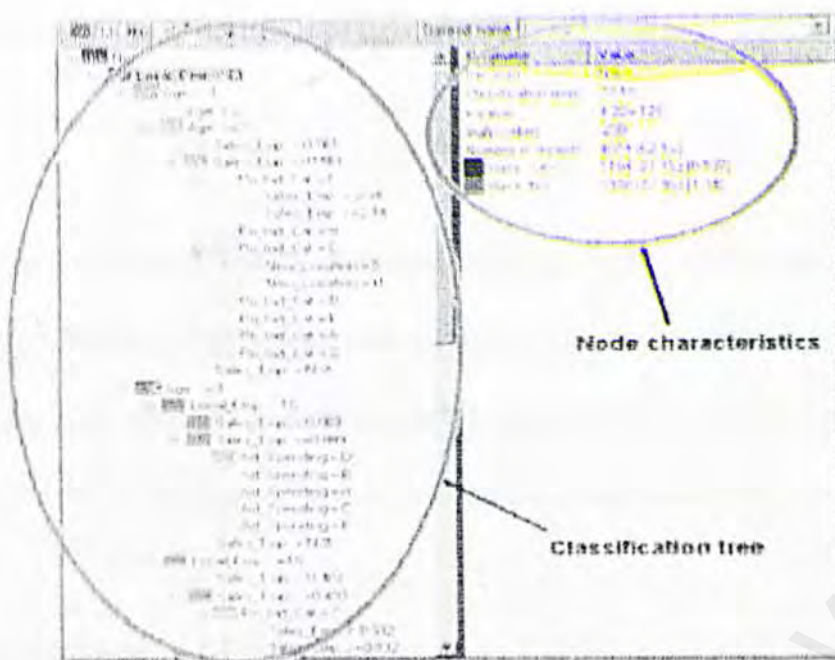


Figure 2.8: Decision Tree Exploration Engine

2.8.2 DBMiner (www.dbminer.com)

DBMiner is a data mining system that originated from the Intelligent Database Systems Research Laboratory, Simon Fraser University, British Columbia, Canada, and has been further developed by DBMiner Technology Inc., British Columbia, Canada.

DBMiner is a cutting edge, on – line analytical mining system running on the Microsoft SQL Server 7.0 Plato system and developed for interactive mining of multiple – level knowledge in large relational databases and data warehouses. The most features of the DBMier system is its tight integration of on – line analytical processing (OLAP) with a wide spectrum of data mining functions, including classification, association, prediction, characterization, and clustering. DBMiner on – line analytical mining (OLAM) servers

AX, SX and DX 2000 are designed to dramatically improve over traditional query and OLAP.

DBMiner integrates smoothly with relational database and data warehouse systems to provide a user – friendly, interactive data mining environment with high performance. This system provides enterprise users powerful capabilities to extract more value and knowledge from their business and operational data to make decisions.

2.8.2.1 The Features of DBMiner

DBMiner offer its latest solutions with several features as show as below: -

- **Power and Accuracy**

DBMiner provides models and scenarios from which users can both understand data relationships and make decisions based on discover knowledge. DBMiner's advanced data visualization functions allow users to present discovered knowledge and results in such comprehensive forms.

- **Easy – To – Use**

While DBMiner offers very sophisticated analytical functions, the underlying data mining technology is well hidden from end users, enabling them to use the software with no knowledge or training in advanced statistical concepts. DBMiner systems are also fully integrated with familiar applications such as Microsoft Excel, with which users can interact to obtain desired results.

- **Openness**

Users have an ability to extract more valuable information from their existing hardware and software system. DBMiner provides an open architecture that is easy to integrate with popular databases and front – end tools, and work with any ODBC, OLE DB and web based sources. DBMiner also enables users to take full advantages of Microsoft Server and its OLAP Services capability.

- **Efficiency and Performance**

An iterative process in which knowledge workers fine – tune their analysis to arrive at a valid conclusion is decisions making. These workers have limited time to wait while their system churns through data for each iteration.

2.8.2.2 Current Products of DBMiner – *DBMiner 2.0*

The DBMiner system has evolved from a research system prototype to a system product. The minimum hardware requirement for the DBMiner system is a Pentium – 550 machine with 64 MB RAM and runs on Windows/NT. There are several products developed by DBMiner, which are DBMiner Insight, DBMiner AX 2002, DBMiner SX 2002, DBMiner DX 2002 and DBMiner 2.0. In this session I just focused to DBMiner 2.0.

DBMiner 2.0 can be freely downloaded at <http://db.cs.sfu.ca/DBMiner> or <http://www.dbminer.com> for 90 – day trial use.

The major new features of DBMiner 2.0 are: -

- A platform change to run over the Microsoft SQLServer 7.0 Plato data warehouse system.
- Addition of a cluster mining module.
- Implementation of a better classification algorithm.
- Use of Excel 2000 to improve the display and flexibility of OLAP functions.

The key features of DBMiner 2.0: -

- **Association** which performing market basket analysis, dependency analysis, and correlation for multidimensional databases
- **Clustering** which that performing data segmentation, grouping data into distinct classes based on the similarity of the properties of the data
- **Classification** which that performing data relevance analysis, market / customer segmentation, and business decision making assistance
- **3D data browsing** which displaying any portions of your data in a 3 – D cube view, and supporting OLAP exploratory analysis on the 3 – D cube
- **Integration with Microsoft SQL Server** which mining directly on MS SQL Server and OLAP data

DBMiner 2.0 Screen Shots: -



Figure 2.9: Decision Tree Classification

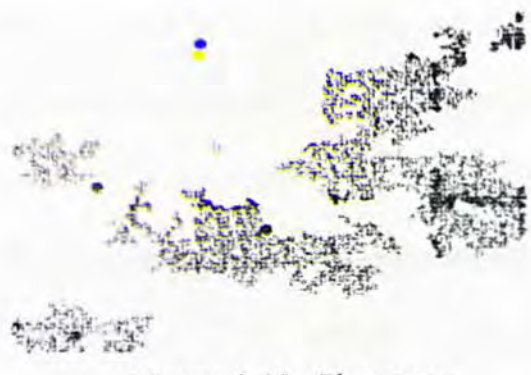


Figure 2.10: Clustering



Figure 2.11: Visual Association

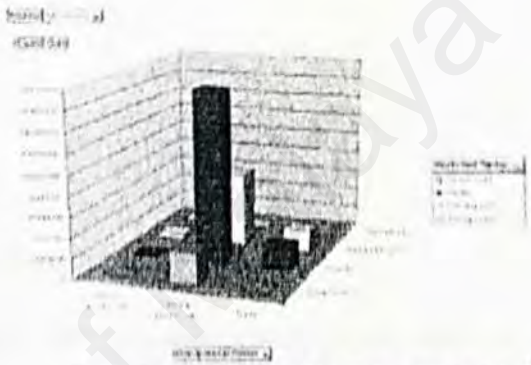


Figure 2.12: OLAP (Summaries)

2.8.3 Clemetine - SPSS Inc. (www.clementine.com)

Clementine, SPSS' Enterprise strength data mining workbench, help businesses improve the profitability of customer relationships through in – depth understanding of data. Organizations use the insight gained from Clementine to retain profitable customers, identify up and cross – selling opportunities, attract new customers, detect fraud, reduce risk and improve government service delivery. Clementine can help businesses better understand their customers'online behavior and improve Web site design, increase e – store sales and deliver online services more effectively.

Clementine 6.0 was the first version to offer a Web mining application template. Whether or not they have any previous experience mining Web data, organizations can use this template to get a jump – start on their Web mining project. Clementine now also includes a data mapping tool to aid users in applying the templates to their own data. In addition to the Web mining application template, Clementine also has a template for use with telecommunications data mining projects. Called “CATS” (Clementine Application Templates), these tools provide end users with collections of pre – built streams, sample data and user guides designed for specific applications.

2.8.3.1 The Features of SPSS

- Capture more datasets with expanded options

SPSS offers many data management features to enhance businesses analysis: -

- Analyze more data faster
- Select cases either permanently or temporarily
- Select random samples or subsets of cases for analysis
- Weight cases by values of a selected variable
- Merge and append files
- Import data from any ODBC – compliant data source, including Microsoft Access

- OLAP report cubes after maximum insight

SPSS' report OLAP (Online Analytical Processing) features give you a fast, flexible way to create, distribute and manipulate information for ad hoc decision

making. Create frequency tables, graphs and report cubes that feature SPSS' unique, award-winning pivoting technology. Add great-looking tables to your reports and presentations that clearly communicate your results. The report cubes are interactive tables that enable you to slice, dice and explore your data like never before. With just a mouseclick, you can explore every angle and aspect of your data.

- Quickly find the answer we need

SPSS is the leading desktop statistical software package, and goes far beyond basic descriptive statistics. You can be confident you'll have the right procedure to get the information you need from your data. The Statistics Coach helps you choose the right procedure for your analysis and the Results Coach helps you understand your results.

SPSS Screen Shots:



name	type	width	decimals	label	values	missing	column
1	1	1	0	1			1
2	1	1	0	2			2
3	1	1	0	3			3
4	1	1	0	4			4
5	1	1	0	5			5
6	1	1	0	6			6
7	1	1	0	7			7
8	1	1	0	8			8
9	1	1	0	9			9
10	1	1	0	10			10
11	1	1	0	11			11
12	1	1	0	12			12
13	1	1	0	13			13
14	1	1	0	14			14
15	1	1	0	15			15
16	1	1	0	16			16
17	1	1	0	17			17
18	1	1	0	18			18
19	1	1	0	19			19
20	1	1	0	20			20
21	1	1	0	21			21
22	1	1	0	22			22

Figure 2.13: SPSS Data Editor

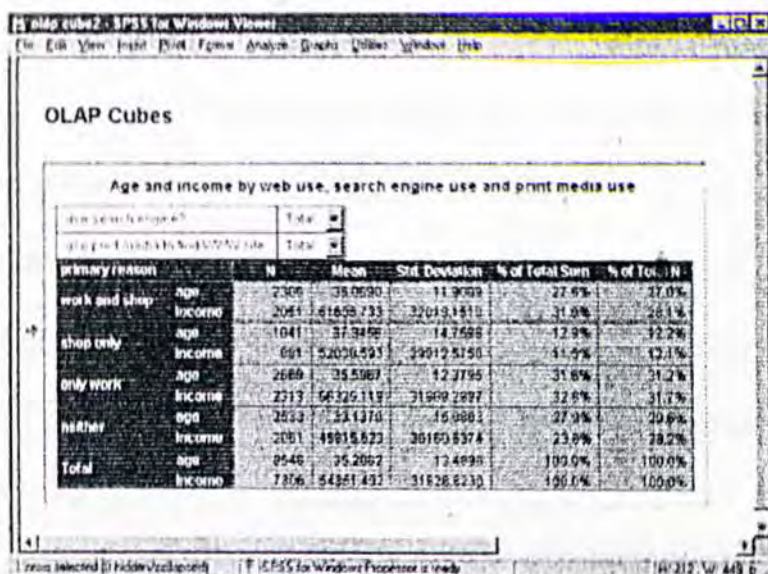


Figure 2.14: SPSS OLAP Cubes

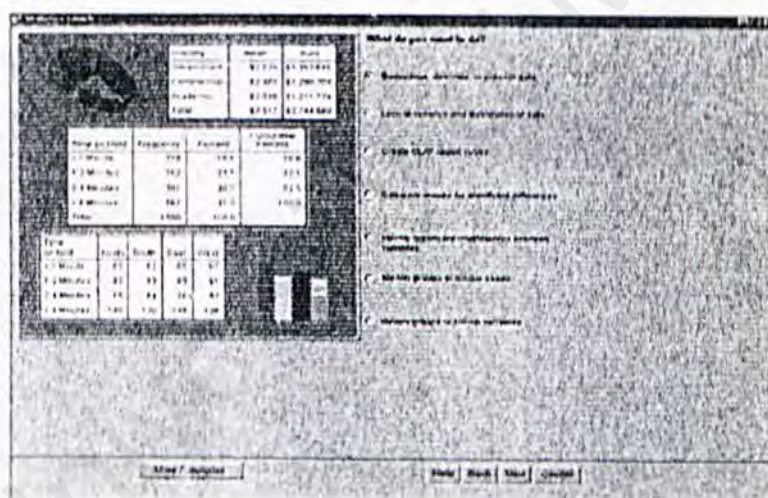


Figure 2.15: SPSS Statistic Coach

2.8.4 IBM DB2 Intelligent Miner for Data

The IBM Intelligent Miner family helps identify and extract high-value business intelligence from your data assets. Through a process of "knowledge discovery," your

organization can leverage hidden information in its data, uncovering associations, patterns, and trends that can lead to real competitive advantage. IBM DB2 Universal Database is the Business Intelligence platform for enterprise wide analytics. IBM Intelligent Miner technologies provide the advanced analytics for you to deliver end – to – end mining solutions. There are 3 types of product of Intelligent DB Miner, which are IBM DB2 Intelligent Miner for Data, IBM Intelligent Miner Scoring, and Intelligent Miner for text. In this part, I just review DB2 Intelligent Miner for Data.

2.8.4.1 IBM DB2 Intelligent Miner for Data

IBM DB2 Intelligent Miner for Data enables users to mine structured data stored in conventional databases or flat files. Its mining algorithms have been successfully used, by customers and partners alike, to address business problems in such areas as customer relationship marketing and fraud and abuse detection. It allows users to increasingly leverage the data warehouse and more quickly derive business value from that investment.

2.8.4.2 Features and Benefits of DB2 Intelligent Miner for Data

- **Single framework** for data mining - a suite of tools to support the iterative process, offering data processing, statistical analysis, and results visualization to complement a variety of mining methods.

- **Proven mining algorithms** that can be used individually or in combination to address a wide range of business problems and deliver measurable business results.
- **Scalable solution** focused on the technical issues of large-scale mining, such as large volumes of data, parallel data mining on AIX, Windows NT, Sun Solaris, and OS/390, directly mining DB2* data, long-running mining operations, and optimization of mining algorithms.
- **Core technology** for IBM data mining solutions, supported by industry-recognized mining consultants deployed worldwide, with customer engagements in finance, telecommunication, insurance, and health care.
- **Application programming interface** enabling development of customized, industry-specific mining applications by customers, IBM, and IBM Business Partners.

2.8.5 Oracle9i Data Mining

Oracle9i Data Mining is an option to Oracle9i Database Enterprise Edition (EE) that embeds data mining functionality for making classifications, predictions, and associations. All model-building and scoring functions are accessible through a Java-based API. Company can build integrated business intelligence applications by using this system. Oracle9i Data Mining helps companies to build business intelligence applications that can find meaningful patterns and associations in corporate data.

2.8.5.1 Competitive Advantages of Oracle9i Data Mining

Oracle9i Data Mining provides several competitive advantages: -

- **Data Mining Embedded in Oracle9i Database**

Oracle9i Data Mining can simplify the process of extracting business intelligence from large amounts of data. All the data mining functionality is embedded in Oracle 9i Database with Oracle9i Data Mining. So the data, data preparation, model building, and model scoring activities all in database.

- **Ability to Enhance Applications with Predictions and Insights**

Oracle9i Data Mining enables companies to systematize the extraction and integration of new business intelligence within their operations. Java – based API of this system can be used by application developers to add data mining insights and predictions to enhance business applications such as CRM, ERP, Web portals, and wireless applications.

- **Java – based API**

Java – based API can be used by application developers to access Oracle9i Data Mining' functionality. Programmatic control of all data mining functions enables automation of data preparation, model building, and model scoring operations.

2.8.5.2 Supported Algorithms in Oracle9i Data Mining

Data Mining algorithms are machine – learning techniques for analyzing for specific categories of problems. There are two algorithms provided by Oracle9i Data Mining,

which are Naïve Bayes for Classifications and Predictions and Association Rules for finding patterns of co-occurring events.

Oracle9i Data Mining Screen Shots:

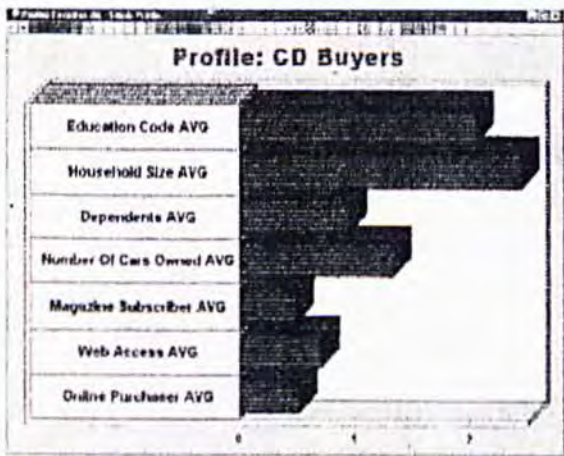


Figure 2.16: Examination of Predictions And Classifications

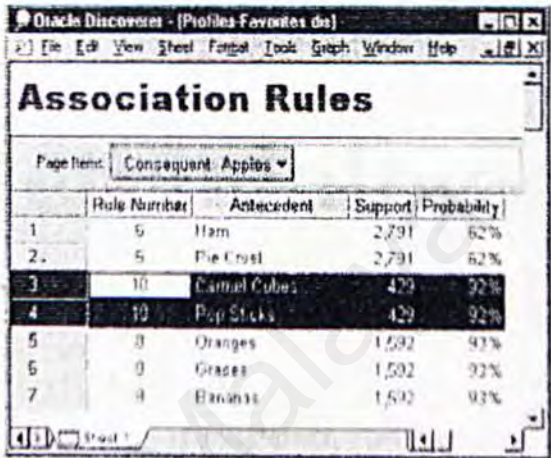


Figure 2.17: Results of Association Analysis

Methodology

- 3.1 *The Waterfall Model*
- 3.2 *Prototyping Model*
- 3.3 *Why Choose Waterfall Model with Prototyping?*
- 3.4 *Data Gathering*

Chapter 3: Methodology

The systems model used for this project is Waterfall Model with Prototyping. This development strategy combines the strengths of both Waterfall Model and Prototyping. The development of a prototype system based on the design is extremely important as it can indicate problems in the original design. It is in this phase that the gathering of important and pertinent information is done.

3.1 The Waterfall Model

The Waterfall Model, which is also called the classic life cycle, or the sequential model is a systematic and sequential approach to software development that begins at the system level and progresses through analysis, design, coding, testing, and support (Pressman, R.S., 1992). With this model, the work begins by establishing requirement for all systems elements.

This system view is essential because software interacts with other elements such as hardware, people, and databases. Thus, it offers a means to make development processes more visible. Because of the cascade from one phase to another, this model is known as the 'Waterfall Model'. So, one phase is completed before the next phase is entered with this model (Bryson J., 1995). The figure in the next page provides a simplified overview of the waterfall method.

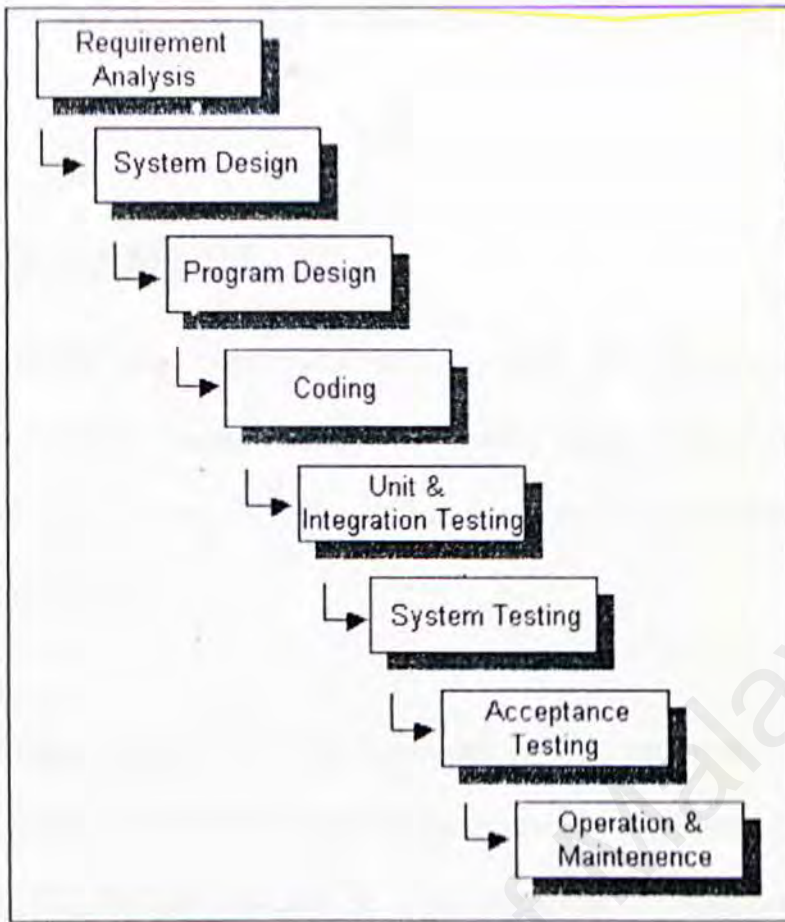


Figure 3.1: The Waterfall Model

The advantage of this model is that once a phase has been completed, it is easy to manage the next phase. The waterfall model can be very useful in helping developers lay out what they need to do. Its simplicity makes it easy to explain to customers who are not familiar with software development; it makes explicit which intermediate products are necessary in order to begin the next stage of development.

The disadvantage is that the project becomes locked into a certain path, and it may be very difficult to deviate. With many computer-assisted learning (and software engineering) projects, design flaws or incorrect specifications are discovered well past

the design phase. Besides that, this model does not reflect the way code is really developed.

3.2 Prototyping Model

Prototyping methods are considered highly useful for developing educational technology. There are a number of different names being used to describe similar design/development methods, including prototyping, rapid application development, rapid prototyping and so on.

Prototyping means creating a partially developed product that enables customers and developers to examine some aspects of the proposed system and decide if it is suitable or appropriate for the finished product. It is an approach for establishing a systems requirements definition that is characterized by high degree of iteration, by a very high degree of user participation in the development process and by an extensive use of approach.

Prototyping provides a communication basis for discussing among all the groups involved in the development process, especially between users and developers. It also enables us to adopt an approach to software construction based on experiment and experience.

The major advantages of prototyping are:

- Able to be created quickly
- Changing the system early in the development

- Relatively inexpensive to built compared to conventional system
- Opportunity to stop development on a system that is not working
- Can determine beforehand appropriateness of design application, the efficiency of computer algorithms, adaptability of operating system and platform in which the system is based
- Serves as risk reduction technique by determining if all aspects of the system are feasible before actual development
- Increases the likelihood that the trial product will be satisfying the needs of the ends users because through interaction the development stages are performed many times

Figure below show the example of the Prototyping Model.

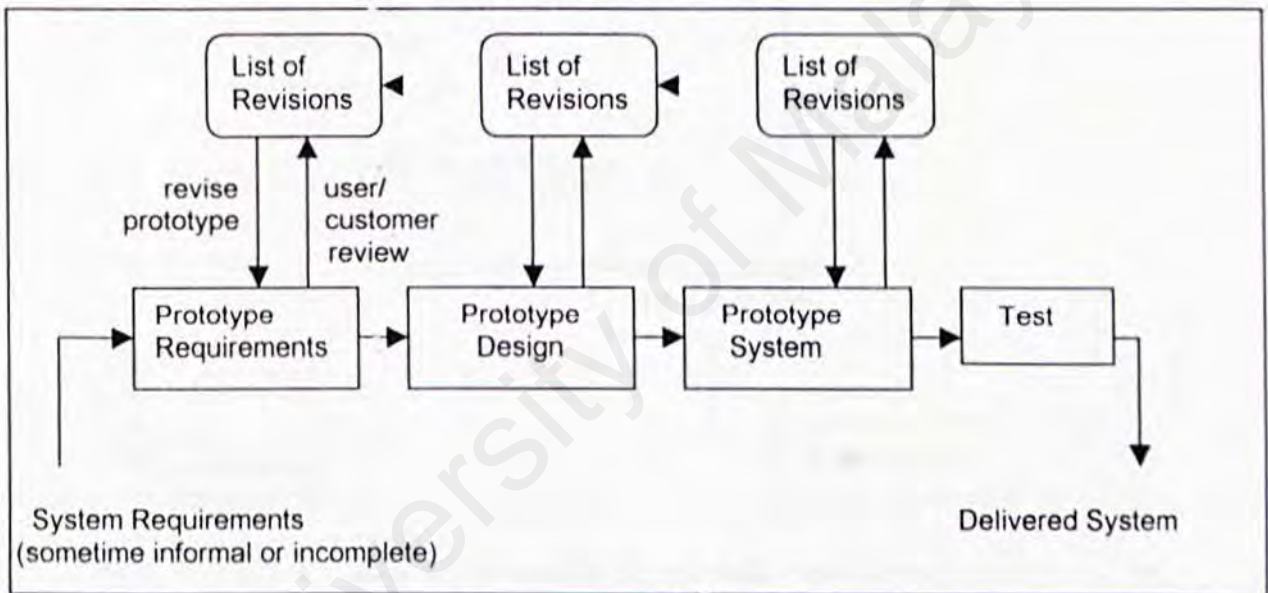


Figure 3.2: The Prototype Model

3.3 Why Choose Waterfall Model with Prototyping?

The combination between Prototyping Model and Waterfall Model will give a better solution for the problems that occur on their own. The prototyping paradigm begins with the requirements gathering. Prototyping of system is a worthwhile technique for quickly

gathering specific information which fits the overall scope and objectives. At the same time, prototyping is involved in the early stages of development where there was a high degree of uncertainty in several areas in the systems requirements. It is used to mainly try out ideas and adapt them appropriately. Figure below shows a model of the Waterfall Model with Prototyping.

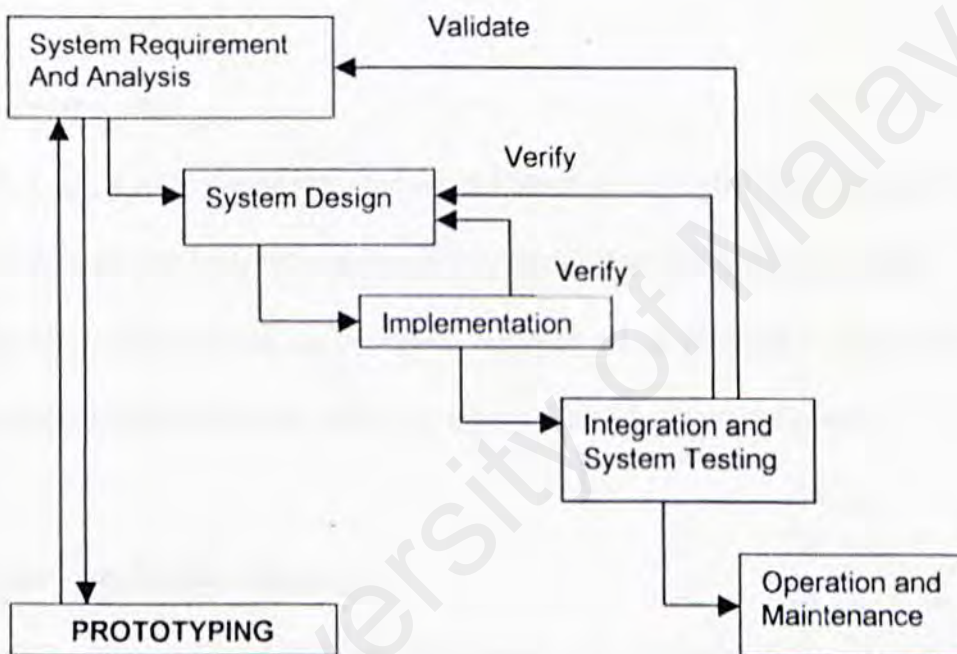


Figure 3.3: Waterfall Model with Prototyping

There are five stages of Waterfall model with Prototyping: -

1. System Requirements and Analysis

The main objective of this stage is to establish the system's services, goals and constraints. All requirements, needs, and constraints in this stage have to identify. All the information regarding this project is gathered where I will discuss in the session

later. The information gathered for this project is through references book from UM library and Internet.

2. System Design

In this stage, draft about data flow diagrams and context diagrams of system models were prepare in the next chapter. The overall system architecture of Data Mining in E-Commerce is established.

3. Implementation

In this stage, a fully functional operational hardware and software systems are produced. Each function will be tested to ensure that the system developed can well – functioning. In this stage also consists of complete, verified set of program components. Coding is the process of translating the design requirements into machine codes.

4. System Testing and Integration

The component of systems are integrated and a properly functioning software is composed in this stage. This integration stage should consider the integration and compatibility of various functions and elements.

5. Operations and Maintenance

This is a stage where updating, deleting, adding and modification of functioning system are done. Maintenance is the act of making adaptations of the software for external changes and internal changes. The three type of maintenance that would be done is corrective measures, adaptive measures, and perfecting measures.

❖ **Doing Researches to Existing System**

In order to get further information and better understanding about data mining system, I have visited to several current data mining existing system such as DBMiner at www.dbminer.com and Oracle⁹ⁱ Data Mining at www.oracle.com and so on. These web sites provided many details of information about their products that are very important for me as a reference tools.

❖ **Internet Surfing**

There are a various type of information could be gathered from Internet nowadays. Despite this, there are many web sites provided useful information for the user in their researches. Some of the web site provided useful software system that could be download or upload from the Internet.

❖ **Brainstorming**

During the requirements elicitation, I always get an advise from my supervisor, Mr. Teh Ying Wah and my friends especially my course mates who also involved in this system development. We always generate lots of ideas as possible to be used in this project.

There are some advantages of Waterfall model with Prototyping that why it's chosen for my project: -

- ✓ It is easy to associate each milestone since each phase is completed before moving on to the next step.
- ✓ It makes the development processes more visible, deliverable and concrete.
- ✓ It enables exploration of alternative strategies and facilitates appropriate revision in order to help identify requirements when there is no current system like the proposed system.

3.4 Data Gathering

To get the data to be used in the various elements of the system, the data gathering phase is very important. Since there is no exact standard or method underlies the process of data gathering, data gathering methods may be vary to suit the needs of each particular project. There are many methods that are used to gather data such as references books, researches, Internet surfing and brainstorming.

❖ References books

UM library is a best place for the student to gathered useful information for their project. UM library has a large collection of journal, references books, magazines, newspaper, example of senior thesis and many more. Students can find the materials they needs with OPAC system, which can make the searching more efficiency. Besides that, document room located in Faculty of Computer Science and Information Technology is another useful place for student to do their research.

4

System Analysis

- 4.1 *Requirement Analysis and Specification*
- 4.2 *System Development Tools Analysis*
- 4.3 *The Tools of Choices*
- 4.4 *System Requirement*

Chapter 4: System Analysis

System analysis is a very essential phase in software development paradigm. This phase is considered the most important phase of all in software development. It comprises of the definition and identification of the problem area, opportunities and objectives. It is also in this phase where all the necessary systems requirements and needs are identified.

4.1 Requirement Analysis and Specification

A requirement is a feature of the system or a description of something the system is capable of doing in order to fulfill the system's purpose. Based on the gathered information, Data Mining in Electronic Commerce requirements have been determined. There are two ways to describe the requirements, which are functional requirements and non – functional requirements.

4.1.1 Functional Requirements Analysis

Requirements describe a system's behavior. Besides that, requirements describe an interaction between the system and its environment (Shari Lawrence Pfleeger, 2001).

The system development in this project is separated to five modules such as Administrator and User Login Module, Data Mining Module, Knowledge Based Module, Print Module and Help Module. Below are the functional requirements respectively.

❖ Administrator and User Login Module

This section will prompt the administrators and users for login name and password as a means of providing security. The administrator login module will only allow administrators to access the system and maintain the database in the system and the user login will only allow user authorized to access the system. This module is vital to protect unauthorized to accessing the system.

❖ Data Mining Module

This module can be separate into 3 sub module, which are: -

- ***Training Set Generator Module***

This module will generate a set of tables from the databases. The result tables are called Training Set which will be used for data mining modeling. Training set is generated by converting field to null, discrete and meaningful value.

- ***Data Mining Model Generator Module***

In this module, data mining algorithms can be generated. For an example, classification algorithm will be used in this project. Decision tree as a classification tool also will be used in the project to classify data into model. Another algorithms are clustering, association and sequential patterns.

- ***Data Visualization Module***

The purpose of this module is to let the user understand what is going on so that they can take action directly. Visualizing a model should allow a user to discuss and explain the logic behind the model with colleagues,

customers, and other users. Decisions about where to put advertising dollars are a direct result of understanding data mining models of customer behavior.

❖ Knowledge Base Module

In this module, data is split into two partitions (old and recent data) so that the partition containing the subset of rows likely to be accessed is smaller.

❖ Print Module

Users can directly print out the information they need such as customers' data, table and graph and so on from this module to help them make the business decisions.

❖ Help Module

Users who have problems when using this system can refer to the user manual in this module as the reference guide to help them use the system easily.

4.1.2 Non – Functional Requirements Analysis

It is important to take non – functional requirements into consideration since it determines the overall quality of the system. A non – functional describes a restrictions on the system that limits our choices for constructing a solution to the problem. The constraints that are important to a system include timing constraints, constraints on the development processes, standards and so on (Sommerville I, 1995)

❖ User – friendly Interface

A standard user interface should integrate the consistent usage of colors, font size, text positioning, graphics and functional menus. An easy – to – use interface will be required for the system as an attractive. There should also use the pop up menus and icons that would guide users to browse through the system. Interface design principles such as user familiarity and consistency should be taken into consideration to develop the interface of this system.

❖ Modularity

The system should be developed in a modular approach to ease maintenance of the system. This approach is important for the decomposition of the modules and functions because it can be broken down into smaller modules.

❖ Robustness

Robustness refers to the quality that causes a system to be able to handle or at least avoid the disaster in the case of unexpected circumstances such as input of improper data. When errors are detected, an error message should be displayed to acknowledge the system user to re – enter data.

❖ Flexibility

The proposed system should be able to accumulate and adapt new technologies and resources as the time and environment changes. This technology includes Object Oriented technology and Advance Security technology. The system should also be adaptable to inside changes and accumulate necessary information.

❖ Reliability & Availability

A reliable system is a system that functions as in program specification and does not produce cost failures when it is use in a reasonable manner. Thus, the system should also be available at all time to allow efficient access to the system.

4.2 System Development Tools Analysis

4.2.1 Operating System

There are two operating system will be analyze here, which are Windows 2000 Professional and Windows NT.

4.2.1.1 Windows 2000 Professional

Windows 2000 Professional is the Windows operating system for business desktop and laptop systems. It is used to run software applications, connect to Internet and intranet sites, and access files, printers, and network resources.

Built on Windows NT® technology and the easy-to-use, familiar Windows® 98 user interface, Windows 2000 Professional gives business users increased flexibility. The integrated Web capabilities let you connect to the Internet from anywhere, at anytime - giving your company access to host of flexible, cost-effective communications options. In addition, broad peripheral and mobile computer support make Windows 2000 Professional an ideal operating system for a workforce that increasingly relies on notebook computers. Further, your support and administrative staff will particularly appreciate the reliability and manageability enhancements that make desktop management simpler and more efficient.

Microsoft® Windows® 2000 features several useful new accessibility tools to help people with disabilities configure and use business computers quickly—without additional software and hardware. Accessibility features from earlier releases of the Windows operating system are still there, and with better compatibility with many assertive technology that now simply work better.

4.2.1.2 Windows NT

Microsoft began development of Windows NT in 1988. Microsoft wished to develop a stable, secure, business-oriented server platform that retained the look and feel of the Windows interface. After five years of development, Windows NT 3.1 was released in 1993. Initially, Windows NT met with little commercial success. Windows NT 3.5 was released in 1994, followed by Windows NT 3.51 in 1995. The security and advanced features (such as long filename support and Internet services) of Windows NT, combined with the Windows desktop, was attractive to businesses looking to replace aging mainframes and terminals.

Unlike Windows 9x, Windows NT was designed to support both CISC (Complex Instruction Set Computers), such as the Intel range; and RISC (Reduced Instruction Set Computers), such as MIPS R4000, Motorola PowerPC, DEC Alpha AXP and Intergraph CLIPPER. However, the support for non-Intel processors has recently been reduced, and at the time of print, the only platforms supported are Intel and DEC Alpha.

As with Windows 9x, networking support was designed into the core operating system at the outset. Previous operating systems such as MS-DOS have network support added later in the form of a series of additional software drivers, known as the shell. A variety of transport protocols and network clients are provided, so that Windows NT can be integrated into complex, mixed network environments, without recourse to any additional software.

4.2.2 Database Management

There are two database management will be analyze here, such as Microsoft Access2000 and Microsoft SQL Server 2000.

4.2.2.1 Microsoft Access 2000

Microsoft Access 2000 is a relational database management system that was created by the Microsoft Corporation for small or home use of storing data in relational data format. It has data access features such as the Remote Data Object (RDO) and Data Access Object (DAO). Thus, Microsoft Access can be used as a database in client/server or n – tier architecture. Microsoft Access provided user – friendly interface and intuitive to create a database easily.

The initial design of database was to be a relational database management system. Then, a web feature was added to Microsoft Access to enable web – based database. The initial Microsoft Access 95 was then updated to Microsoft Access 2000 with some more features.

Microsoft Access provides a very easy, quick and convenient channel to publish both the static and dynamic data to the web. Its features of software are very common and is

widely used and owned due to its ease of use. It also provides a cheap mean to test the web database publishing extremely well.

4.2.2.2 Microsoft SQL Server 2000

Microsoft SQL Server is a significant tool in many regards, from data warehousing to application that require not only a large amount of information but also many different simultaneous users. It is also a key component in answering data management requirements and a powerful as well as comprehensive database.

Microsoft SQL Server is the compact database for rapidly developing applications that extend enterprise data management capabilities to devices. It is also a perfect example of an n-tier system. The users can manipulate the data directly from the client side. Most of the time, the data is validated first before it is updated into the database in server side.

It is tightly integrated with Microsoft BackOffice family products as well to enable organization to improve decision making and streamline the business process. And it is admittedly the best database for Windows NT Server 4.0.

4.2.3 Application Programming Language

There are two programming language I will discuss here, which are JAVA and Visual Basic 6.0.

4.2.3.1 JAVA

Java is a perfect programming language for building applications for the Internet. It's a programming language that used to write programs called applets that can be run from within Web pages. Java programs can also be designed as standalone application, which they run on their own, independent of Web pages. Java is a platform – independent that can be run on just about any computer (Dr. Harvey M.Deitel & Paul J. Deitel, 1999).

Java is a relatively new programming language. There are several reasons why this language has appeal: -

- i. Java is largely platform independent, means that an application written for one computer is very likely to run unchanged on another computer. Thus, a single application can be written to execute across all of a company's computers, whether they are PC's, Macs, or Unix workstations.
- ii. Java has a C – like syntax that means it is partially familiar to millions of people.
- iii. Java is object oriented, which should make code written in Java more reusable between applications
- iv. Java is possible to write device – independent graphics applications
- v. Java is free. Sun provides a free Java Development Kit for download from its WWW site (<http://java.sun.com>).

Java differs from other computer language in that all Java programs are compiled to execute on a special computer known as the Java Virtual Machine (Java VM).

4.2.3.2 Microsoft Visual Basic 6.0

Visual Basic is an extremely powerful, full – featured application development tool that exploits the key features of Microsoft Windows. It is easy – to – use through a graphical interface where applications can be built in the short time by using it.

Advanced database applications can be developed to access SQL Server database, Access database or any third – party database by using ODBC, DAO, RDO, or ADO and bind the data to forms and reports that greatly reduce development time by using integrated visual database tools (Dr. Harvey M.Deitel & Paul J. Deitel, 1999).

Visual Basic 6.0 provides support for Graphical User Interface (GUI) design that helps interface designer to enhance screen design. Besides that, Visual Basic 6.0 is also an integrated language system that enables users to test and debug application on the fly from within the development environment.

4.2.4 Application Programming Software

There are three types of programming software that I will discuss which are JBuilder 7.0, JCreator 2.5 and TextPad.

4.2.4.1 JBuilder 7.0

JBuilder is the most comprehensive award-winning visual development environment for building applications, applets, JSP/Servlets, JavaBeans, Enterprise JavaBeans and distributed J2EE applications for the Java 2 Platform. JBuilder 7 is available for Windows, Linux, Solaris and Mac OS X. JBuilder 7 contains major improvements in

developer productivity, as well as a cleaner, more intuitive user interface and dramatic performance enhancements.

JBuilder 7.0 support for the latest Java™ Standards. The Borland® JBuilder™ environment is hosted on the JAVA 2 SDK 1.3 and is entirely implemented in Java for excellent platform interoperability and performance on Windows, Linux, Solaris, Mac, OS, and any operating system that fully support the JAVA 2 SDK 1.3.

JBuilder combines a unique set of features that allows for a fast, interactive development life cycle. Configuration management supports complex build processes. Besides that, JBuilder uses tools for performing the presentation, transformation, and validation of XML docs.

4.2.4.2 JCreator 2.5

Creator is a powerful Free IDE for Java technologies. JCreator provides the user with a wide range of functionality such as: Project management, Project Templates, Class browsers, syntax highlighting, wizards and a fully customizable user interface. Below are some features of JCreator 2.5: -

- Selection margin with line numbers

JCreator gives you the option of viewing line numbers in the selection margin.

- Instant colour syntax highlighting

With syntax highlighting, you can easily distinguish between keywords, methods, comments and plain text. All colors can be customized to your preferences. The syntax parser works with syntax configuration files which you

can modify to your needs. The parser handles up to four keyword lists and one list with preprocessors.

- Unlimited Undo/Redo capability
- Visible display of tabs and spaces
- A powerful search and replace engine for single and multiple documents

The result of the search will be displayed in the output view.

- Friendly user – interface
- Code Templates

The Code Templates are abbreviations you may expand to strings while editing the code. They are expanded in the text by entering the abbreviations and pressing either Space, Enter or Tab. You can also trigger the code template popup-list using the shortcut defined in the edit menu.

- **Implement Interface Wizard**

This tool inserts code for implementing Interfaces.

4.2.4.3 TextPad

TextPad is a versatile text editor that has a myriad of features. TextPad can edit multiple files, transpose characters or lines; supports unlimited undo/redo, a spelling checker with dictionaries in 13 languages, sorting up to 3 keys, block selection mode, a keystroke macro recorder, and syntax highlighting for a variety of languages including ASP, C/C++, HTML, Java, Perl, and PHP. TextPad has a powerful search/replace engine as well, which not only supports the use of UNIX-style regular expressions, but allows search/replace across multiple documents.

Starting with version 4, TextPad implements a core functionality called Document Classes that lets you define different settings for different families (or classes) of files. With this feature, for example you can make sure that HTML files will always be saved in Unix format while text files will always be displayed with word wrapping on. TextPad also has the most easy-to-use Preferences dialog box we have ever seen in a text editor thanks to its two-pane tree-view interface.

Perhaps, the most impressive feature of TextPad is its bullet-proof unlimited undo feature. TextPad allows you to revert to the original version of the document no matter what you did even after saving it several times (just uncheck the Clean undo buffer before saving in Configure > Preferences > Editor).

TextPad is designed to provide the power and functionality to satisfy the most demanding text editing requirements. The 32-bit edition can edit files up to the limits of virtual memory, and will work with MS Windows™ 9x, Windows NT and Windows 2000.

4.3 The Tools of Choices

4.3.1 Operating System - Windows 2000 Professional

Windows 2000 Professional is the Windows operating system for business desktop and laptop systems. It is used to run software applications, connect to Internet and intranet sites, and access files, printers, and network resources.

The reason why I choose Windows 2000 Professional as my application platforms for this project are: -

- **Work how and where I want** with new peripheral support and features that extend notebook capabilities.
- **Rely on my PC** to be up and running with enterprise level quality.
- **Work the way I did with Windows 98, only much faster.** Combine the ease of Windows 98 with the manageability, reliability, and security of Windows NT, at speeds 30% faster than Windows 98 on PCs with 64 MB of RAM or more.
- **Communicate, share information, and use the Internet quickly and easily.** With integrated support for Internet – enabled applications, business software developers incorporate the new ways to create and share information made possible by the Internet.

4.3.2 Database Management – Microsoft Access 2000

Microsoft Access 2002 offers a set of tools rich enough for the most experienced database developer, yet accessible enough for first-time users.

The reasons why choose Microsoft Access 2000 in this purposed project are: -

- It allows generating, analyzing and creating report quickly
- It can be integrated easily Microsoft Front Page 2000
- It integrates data from spreadsheets and other databases, and thus provides an easy way to find answers and share information over intranets and the internet

- It has many simple and user – friendly features too build tables, queries, forms, and reports that can be further customized to suit project need.
- Data Access Pages allows users to extend database applications to the corporate intranet by creating data – bound HTML pages quickly and easily.
- A toolbox is provided in the Data Access Page design environment for creating controls. Users can easily drag-and-drop each tool as needed.
- Control Grouping allows users to group controls as a single unit to make form design easier.

4.3.3 Application Programming Language – JAVA

JAVA was chosen as this project's programming technology because of this is a high – performance language and it make the language well suited for use on the World Wide Web. Java technology – based software can work everywhere.

The reasons why I choose Java as my programming language are: -

- Java is largely platform independent, means that an application written for one computer is very likely to run unchanged on another computer. Thus, a single application can be written to execute across all of a company's computers, whether they are PC's, Macs, or Unix workstations.
- Java has a C - like syntax that means it is partially familiar to millions of people.
- Java is object oriented, which should make code written in Java more reusable between applications.
- Java is possible to write device – independent graphics applications

4.3.4 Application Programming Software –TextPad

TextPad was chosen as an addition for this project because of its power and functionality to satisfy the most demanding text editing requirements.

The reasons why I choose TextPad as my programming software are: -

- Huge files can be edited.
- Supports Universal Naming Convention (UNC) style names, and long file names with spaces.
- CUA compliant keyboard commands, which can be fully customized.
- Customizable color syntax highlighting.
- Multiple personalities: compatibility with Microsoft applications, WordStar, BRIEF and TextPad 2.
- The user interface is available in French, German, Italian, Japanese, Polish, Portuguese, Russian and Spanish.
- Multiple files can be simultaneously edited, with up to 4 views per file.
- The multiple workplace feature lets you restart exactly where you left off.

4.4 System Requirement

The choice of hardware and software requirements is important in developing a system.

4.4.1 Hardware Requirement

Table below is the minimal hardware requirement for developing my project.

Component	Descriptions
Processor	- 333 MHz or higher Pentium compatible
RAM	- At least 64 MB memory
Hard disk	- At least 2 GB of hard disk
Other standard input / output devices.	

Table 4.1: Hardware Requirement

4.4.2 Software Requirement

Table below is the software requirement for developing my project.

Software	Descriptions
Operating System	- Windows 2000 Professional
Database Management	- Microsoft Access 2000
Programming Language	- Java
Application Programming Software	- TextPad

Table 4.2: Software Requirement

Chapter

5

System Design

- 5.1 Architectural Design of the Project*
- 5.2 System Structure Chart*
- 5.3 Data Flow Diagrams (DFD)*
- 5.4 Entity Relationship (ER) Diagram*
- 5.5 Database Design*
- 5.6 User Interface Design*

Chapter 5: System Design

The phase of system design is another important phase that contributes towards a good system. In this phase, the requirements gathered from the previous phase, which is system analysis. The system design for Data Mining in E - Commerce using Classification Technique is divided into the following stages:

- System Structure Chart
- Database Design
- User Interface Design

5.1 Architectural Design of the Project

Any system that is to be designed must go through the decomposition. Decomposition is the process of breaking down the larger program into sub – systems that are more understandable. The initial process of this phase is to identify these subsystems and then establish a framework that can simplify the sub – system's control and communication. This is what architectural design is all about.

The first phase of architectural design will be on decomposing the system into sets of interacting component of subsystems. Next, communications between the subsystems are identified. At its most abstract level, an architectural design is depicted as block diagram to represent an overview the system structure and to describe the interactions that exist between the many independent subsystems.

5.2 System Structure Chart

System structure charts are used to depict high-level abstraction of a specified system. The use of a structure chart is to describe the interaction between independent modules. Major functions from the initial component part of the structure chart, which can be broken into detailed sub-components. (Jeffrey L. Whitten, Lonnie D. Bentley & Kevin C. Dittman, 2000). Figure 5.1 below will illustrate the whole structure chart of Data Mining in Electronic Commerce using Classification Technique for each module. This system is divided into two modules and sub – modules.

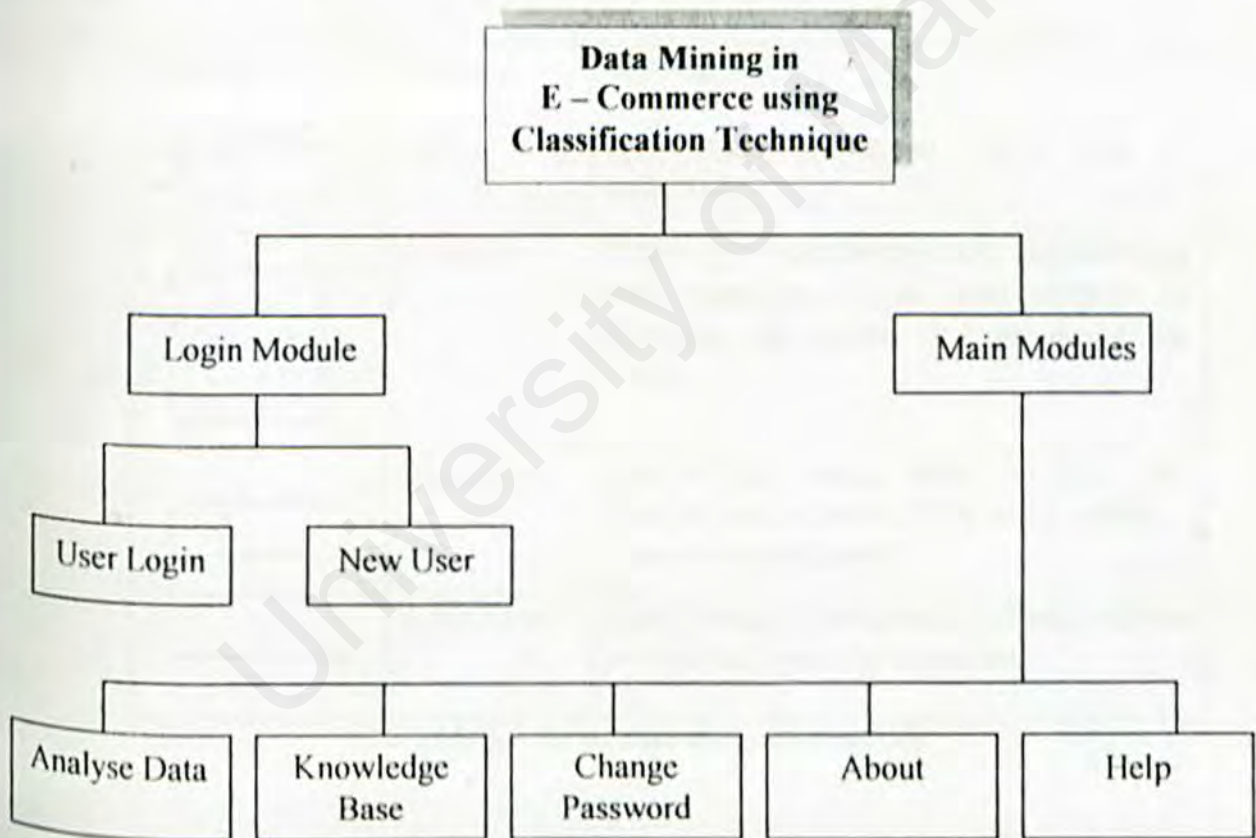


Figure 5.1: System Structure Chart

5.3 Data Flow Diagrams (DFD)

A data flow diagram (DFD) is a graphical system model that shows all of the main requirements for an information system in one diagram: inputs and outputs, processes, and data storage. Everyone working on a development project can see all aspects of the system working together at once with the DFD. That is one reason for its popularity. The DFD is also easy to read because it is graphical model and because there are only four symbols to learn. The DFD provides the system analyst with the ability to specify a system at the logical level (what the system does) rather than the physical level (how it does).

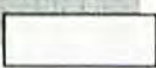

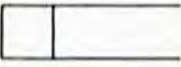

Symbol	Name	Description
	Entity	Any object or event which data is collected
	Process	Step – by – step instructions are followed that transform inputs into outputs (a computer or person or both doing the work).
	Data Store	Data at rest, being stored for later use. Usually corresponds to a data entity – relationship diagram.
	Data Flow	Data flowing from place to place, such as an input or output to a process.

Table 5.1: Data flow diagram symbols

5.3.1 Context Diagram

A context diagram is a DFD that describes the highest – level view of a system. All external agents and all data flows into and out of the system are shown in one diagram, with the whole system represented as one process.

Below is the context diagram for the Data Mining in Electronic Commerce:

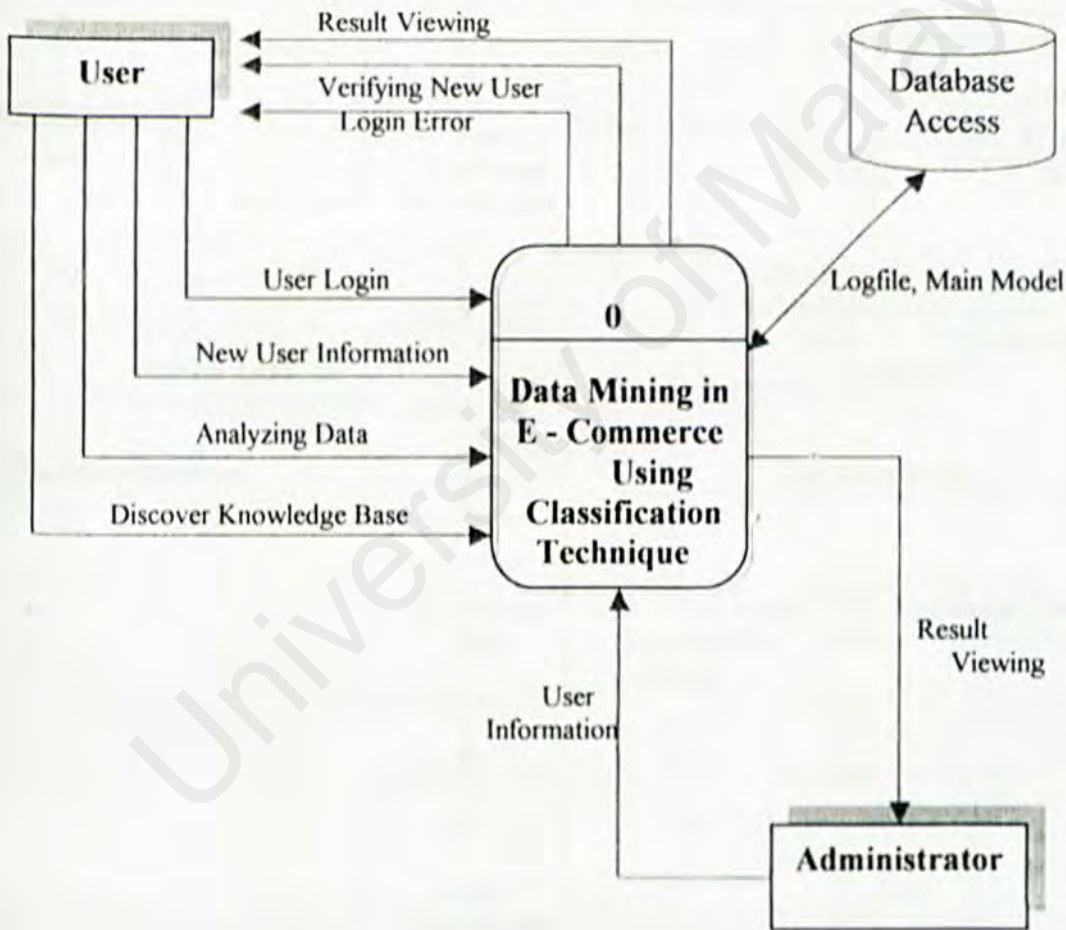


Figure 5.2: Context Diagram for Data Mining in E-Commerce using Classification Technique

5.3.2 Diagram 0 (Level 0)

Diagram 0 is used to show the entire system on a single DFD in greater detail than on the context diagram. Each process on diagram 0 represents processing for a single event. It also shows all processes and data stores involve in Data Mining in E-Commerce using Classification Technique. A figure 5.3 shows the diagram 0 for this system.

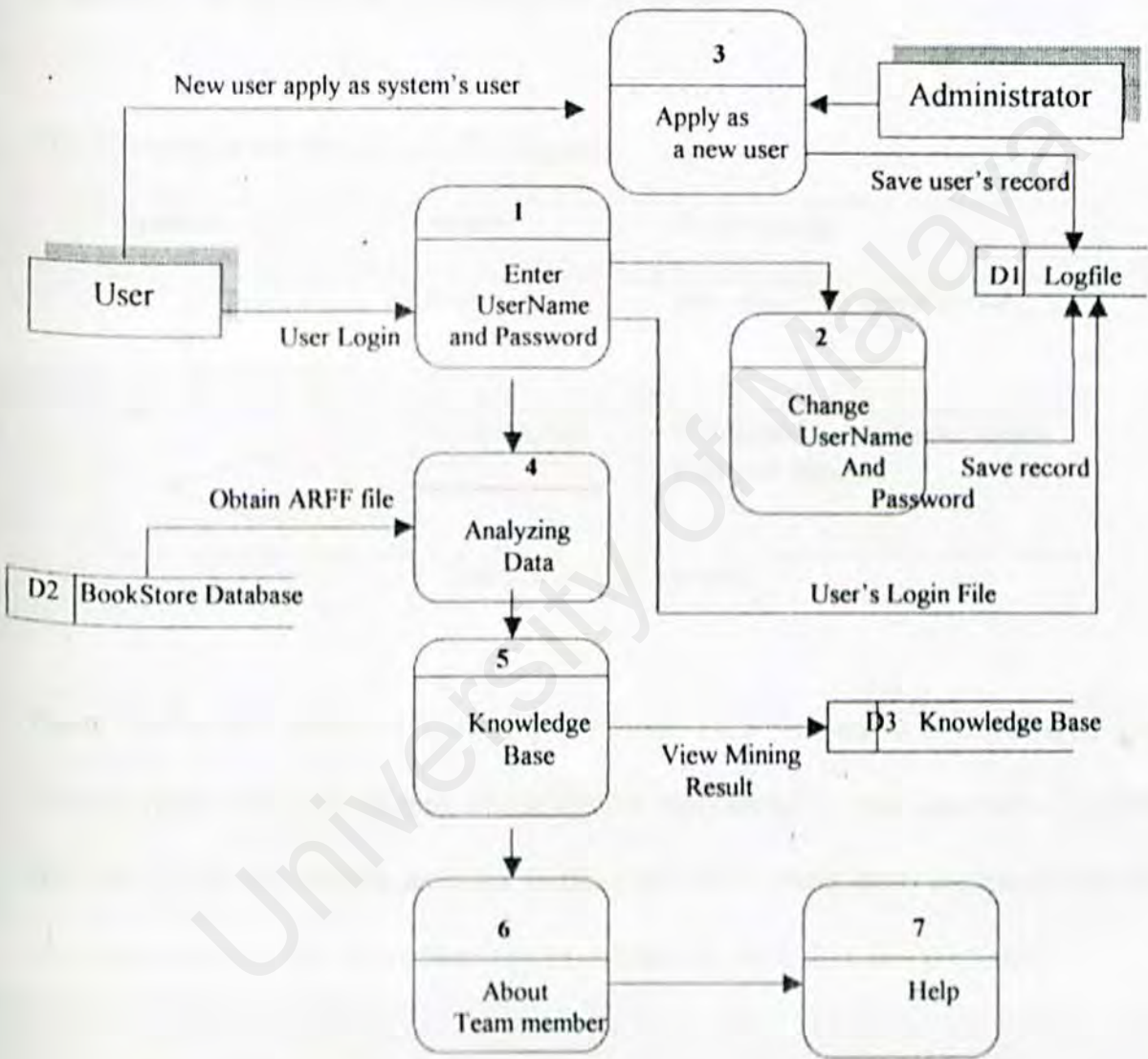


Figure 5.3: Diagram 0 for Data Mining in E-Commerce

5.4 Entity Relationship (ER) Diagram

The Entity – Relationship (ER) Data Model is a detailed, logical representation of the data for an organization or for a business application. The model is expressed in terms of entities in the business environment, the relationships between entities, and the attributes of both entities and relationships. The ERD (Entity – Relationship Diagram) is used to graphically represent an ER Data Model (P. Sellappan, 2000).

The following is the symbol of E-R diagram.



Symbol	Name	Description
	Entity	Any object or event about
	Relationship	Relationships are association between entities

Table 5.2: ERD symbols

Figure 5.4 below shows the ERD for the Book Data Mining in E-Commerce using Classification technique. Despite of this system was created by two databases, Customer and Item, it is a very simple diagram. In this case, M:N means many customer can order more than one Item and many Item can be ordered by more than one customer.



Figure 5.4: ERD for Data Mining in E-Commerce using Classification Technique

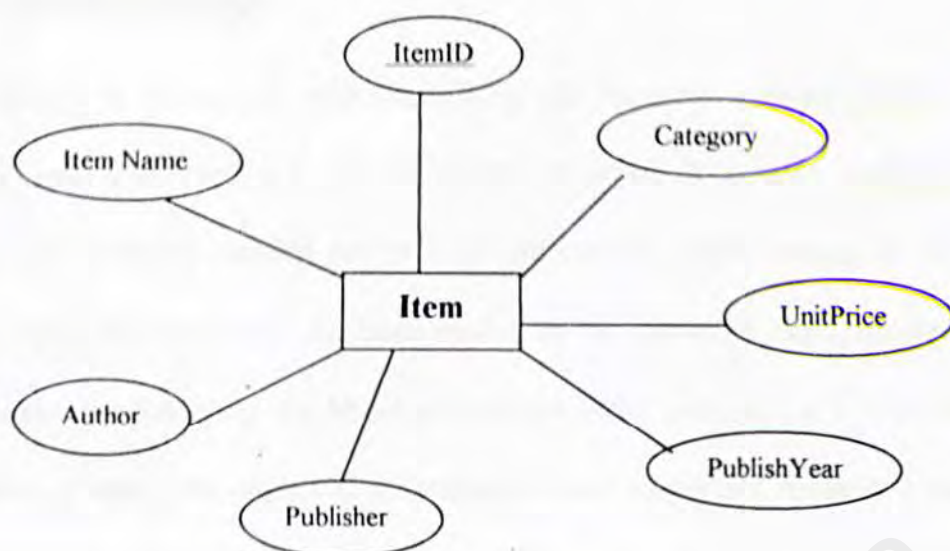


Figure 5.5: Item Entity and its attributes

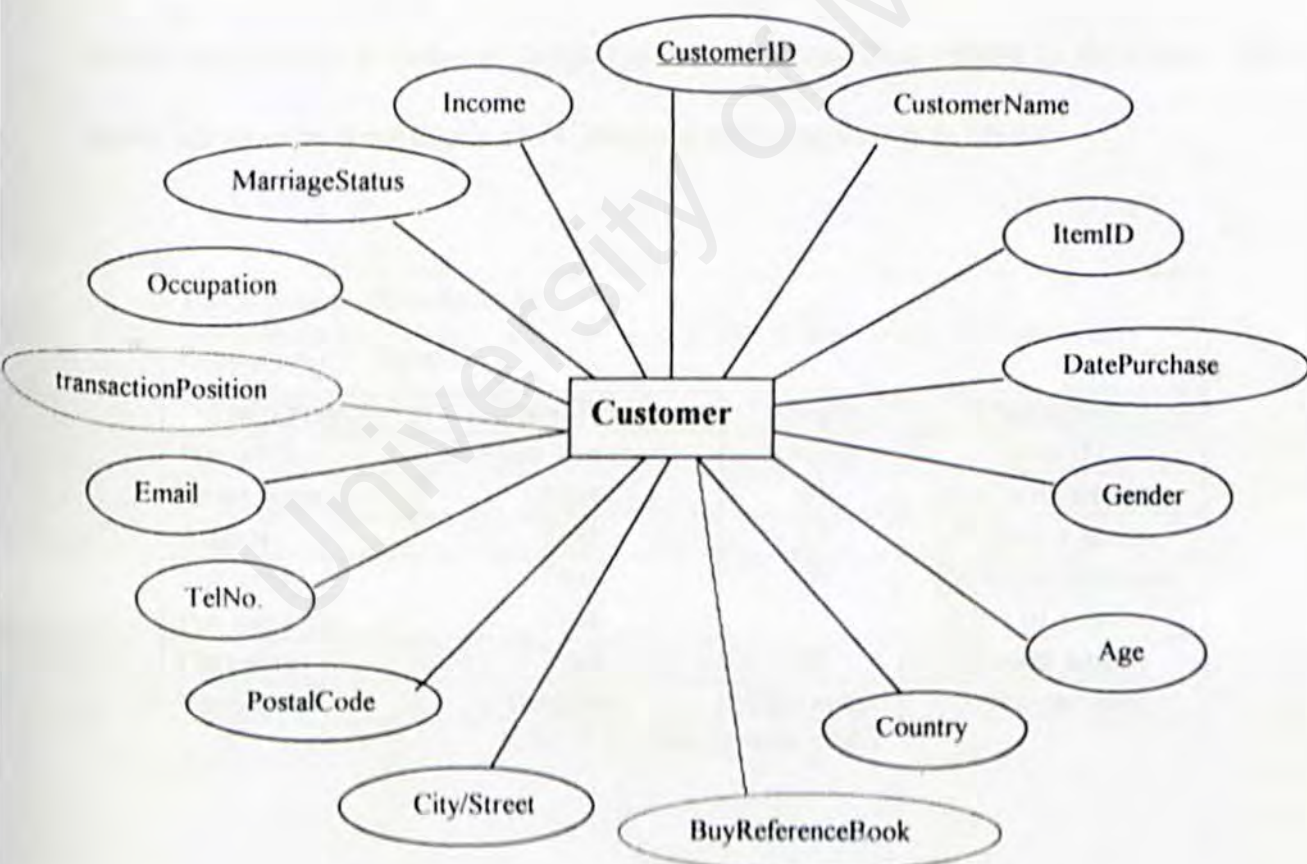


Figure 5.6: Customer Entity and its Attributes

5.5 Database Design

Database design is concerned with identifying the business entities relevant to the application, and how they are related to one another. It is also concerned with identifying the attributes needed for each of the entities. *Data Mining in Electronic Commerce* uses the relational database model in its database implementation. The database is constructed using the Microsoft Access 2000. Database is a collection of a large amount of data. The design of a database is very important because it can affect greatly the performance of data retrieval, updating and query as well as in the runtime period of the system.

The database for Data Mining in E-commerce using Classification Technique is divided into 2 main tables in order to keeps the record or any data related to the system. The tables are namely ItemDetails and CustomerDetails as shown as below.

Table Name: ItemDetails			
Primary Key: ItemID			
Field Name	Data Type	Length	Description
*ItemID	Auto Number	Integer	Item ID
ItemName	Text	50	Item's name title
Author	Text	50	Author's name
Publisher	Text	60	Publisher's name
PublishYear	Text	10	Year of publish
Category	Text	50	Group name
UnitPrice	Currency	Currency	Price per unit

Table 5.3: ItemDetails Table

Table Name: CustomerDetails			
Primary Key: CusID			
Field Name	Data Type	Length	Description
*CustomerID	Text	10	Customer ID
CustomerName	Text	50	Customer's name
ItemID	Auto Number	Integer	Item ID
Gender	Text	10	Customer Gender
Age	Text	20	Customer Age
City/Street	Text	100	City of address
Postal Code	Text	10	Postal code of address
Country	Text	25	Country of address
Tel	Text	20	Phone Number
Email	Text	100	Email address
Occupation	Text	50	Customer's occupation
MarriageStatus	Text	20	Customer's marriage
Income	Text	50	Customer's income
transactionPosition	Text	20	Customer's transaction position
DatePurchase	Date/Time	ShortDate	Date of purchase
BuyReferenceBook	Text	50	Class Label

Table 5.4: CustomerDetails Table

5.6 User Interface Design

A good user interface allows people to work with the application or system without having to read the manuals or receiving training. Interface design is important for several reasons. First of all, the more intuitive the user interface, the easier it is to train people to use it, to reduce the training costs (Kendall K.E & Kendall J.E., 1999).

User interface of a system is always used as a measure or yardstick by which the system is judged. Any interface that is hard to use or understand will result in higher level of

user errors. This is because the level of understanding of each user is different. It might lead to software system to discard, irrespective of its functionality, at worst.

The following shows few of the interface design templates for the purposed Data Mining in E – Commerce using Classification Technique.

Login

User Name :

Password :

Authorization

Login

Cancel

New User

Figure 5.7: Login Module Interface

Add New User

User Name :

Password :

Confirm Password :

Status :

OK

Cancel

Figure 5.8: Add New User Interface

Data Mining in E - Commerce

Analyse Data
Knowledge Base
Change Password
About
Help
Logout

Figure 5.9: Main Module Interface

Analyse Data

Association	Classification	Clustering	Sequential
-------------	----------------	------------	------------

Classification Setting

Cancel	Reset	Calculate	More
--------	-------	-----------	------

Figure 5.10: Analyse Data Interface

Chapter

6

System Implementation

- 6.1 Development Environment*
- 6.2 Program Development*
- 6.3 Program Coding*
- 6.4 Database Implementation*

Chapter 6: System Implementation

System implementation is well defined as the integration of the physical, conceptual and constructed resources that produce a working system. Therefore, system implementation is the physical realization of the database and application designs. In system implementation, database will be created and source code will be written, which will produce and deliver a functional system (Roger S. Pressman, 2000).

The system implementation of the Data Mining in Electronic Commerce using Classification Technique will be divided into three components, which are the development environment, program development and program coding.

6.1 Development Environment

6.1.1 Hardware Configuration

Data Mining in Electronic Commerce using Classification Technique was developed using a computer with hardware specification as described in the table below: -

Hardware Component	Specifications
Central Processor Unit	Intel PentiumIII 333 MHz
Memory	80MB
Cache Memory	512K Pipeline Burst Cache
Hard Disk	2 GB
CD – ROM Drive	6 x speed
Floppy Drive	1.44MB
Monitor	15" SVGA

Other standard desktop PC components

Table 6.1: Hardware specification of computer used for system development

6.1.2 Software Tools

6.1.2.1 Software Tools for Development

A listing of software used in the development of Data Mining in Electronic Commerce using Classification Technique is provided in the table below: -

Software	Phase/Process
Windows 2000 Professional	Operating System
TextPad	Programming tools in Java
Microsoft Access 2000	Creating databases

Table 6.2: Software tools for Development

6.1.2.2 Software Tools for Report Writing

Microsoft Word 2000 is used to write the report because of its wide availability and user friendliness. Microsoft Word 2000 was used to write the report and draw the Structure Chart, DFD, ER Diagram, and System Model. It was also used in report writing and creating user manual which comprises the documentation part of this project.

6.2 Program Development

Program development is the process of creating the programs needed to satisfy an information system's processing requirements. There are 5 steps in the program development, which are review the program documentation, design the program, code the program, and test the program and completion the program documentation as shown as in Figure 6.1 next pages.

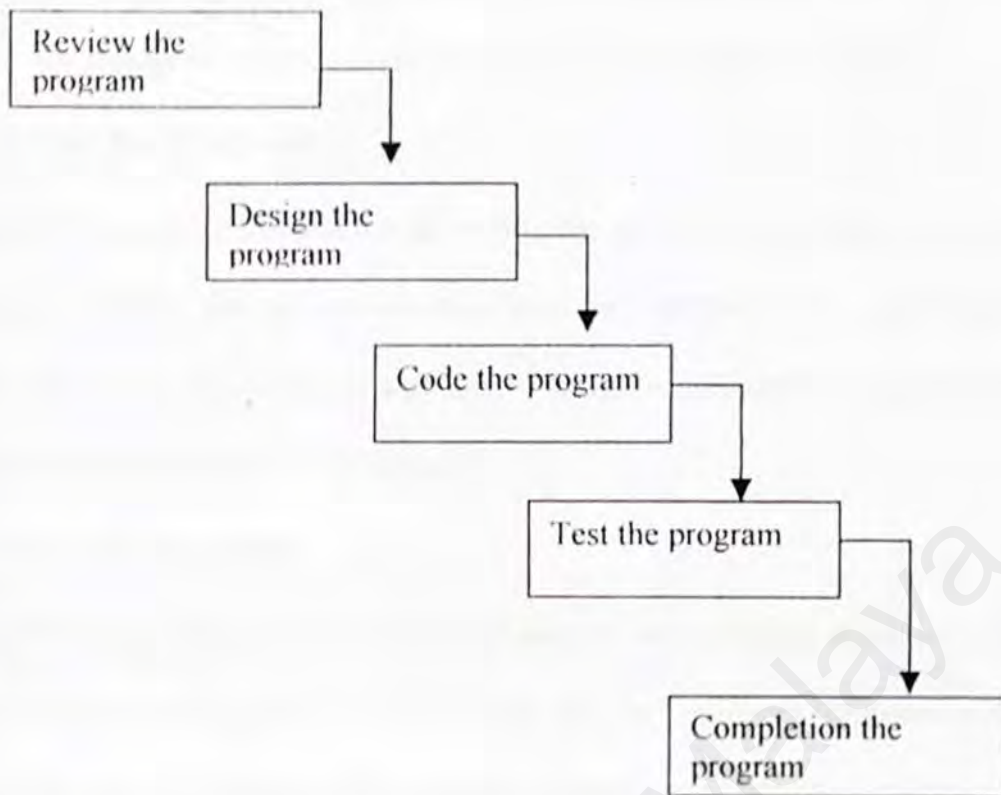


Figure 6.1: The five steps of Program Development

6.2.1 Review the Program Documentation

The first step in the program development is to review the program documentation that was prepared during the previous phases. The program documentation of catalogue ordering system consists of simple process descriptions, report layouts, data dictionary entries and the source documents. This documentation helps me to understand better the work that needs to be covered during this coding phase.

6.2.2 Design the program

After the program documentation review, I need to design the program, which is the second level of program design during the system development. For this second level of program design, I have exactly decided how the program can accomplish what it must

do by developing a logical solution to the programming problems. The logical solution, or logic, for a program is a step-by-step solution to a programming problem.

6.2.3 Code the Program

Coding the program is the process of writing the program instructions that implement the program design. Design specification must be translated into a machine-readable format. The coding step performs this task. If design is performed in a detailed manner, coding can be accomplished mechanically.

6.2.4 Test the program

During the testing program level, I must thoroughly test a program to ensure it functions correctly before the program processes actual data and produces information on which people will rely. I will perform several types of test on an individual program. (will be further discusses in details in section below).

6.2.5 Document the program

Accurate and complete program documentation is essential for the successful operations and maintenance of the information system. This documentation includes the system user manual that may need by most of the customers as well as the system administrator's.

6.3 Program Coding

Coding is a process that translate a detail design representation of software into a programming language realization (R.S.Pressman, 1992)

6.3.1 Methodology

Data Mining in E – Commerce using Classification Technique is developed using a modular approach where each module is developed separately and are later integrated into a fully functional system. For each module, it is further refined into functions and procedures. By using a modular approach, future modification and enhancements are made easily.

6.3.2 Coding Principles

The following principles were applied during the implementation of Data Mining in E – Commerce using Classification Technique:

➤ Coding Conventions

Coding conventions such as program labeling, naming conventions, comments and indentation should be adhered to. It provides easy identification for the programmer.

➤ Readability

Codes should be easy to understand. Adherence to coding conventions such as naming conventions and indentation contribute to program readability.

➤ **Maintainability**

Codes should be easily revised or corrected. To facilitate maintenance, code should be readable, modular and as general as possible.

➤ **Robustness**

The codes should be able to handle cases of user error by responding appropriately.

➤ **Internal Documentation**

Internal comments provide a clear guide during the maintenance phase of the system. Comments provide the developer with a means of communicating with other readers of the source code. Statements of purpose indicating the function of the module and a descriptive comment that is embedded within the body of the source code is needed to describe processing functions.

6.3.3 Database Connectivity

In order to connect to the database, Open Database Connectivity (ODBC) was created in the server by specifying the Data Source Name (DSN). ODBC is an application programming interface (API) for database access. By using ODBC statements in a program, I can access data in Microsoft Access.

JDBC – ODBC Bridge also been used to reach ODBC – accessible databases. Java Database Connectivity (JDBC) makes migration easier. Besides that, SQL statements had been used in this project. SQL allows users to access data in relational databases in MS Access. Below shows the example of database connectivity for the project.

```

/***** Get database connection *****/

public Connection getDBCConnection()
{
    //Set up connection

    String url = "jdbc:odbc:BookStore";

    Connection conn = null;

    try
    {
        Class.forName("sun.jdbc.odbc.JdbcOdbcDriver");

        conn = DriverManager.getConnection(url, "", "");

    }

    catch ( ClassNotFoundException cnfex)
    {
        System.err.println("Fail to load Jdbc/Odbc driver.");

        cnfex.printStackTrace();

        System.exit(1);

    }

    catch (SQLException sqlex )
    {
        System.err.println("Unable to connect");

        sqlex.printStackTrace();

    }

    catch(Exception ex)
    {
        System.out.println(ex);

    }

    return conn;
}

```


6.4 Database Implementation

Microsoft Access 2000 was used to create and manage the database for storing data from Data Mining in E – Commerce using Classification Technique (BookStore). MS Access 2000 was used because of its wide range of functionalities and ease of use for fast database development.

Microsoft Access 2000 offers improved 32 – bit performance, including smaller forms, more compilation and better data manipulation technology that result in quicker responses and faster data operations. The Performance Analyzer looks at the database and suggests ways to speed it up.

Access 2000 has an integrated development with Microsoft Visual Basic for Applications and ActiveX. Its Intuitive IDE features include drag – and drop code, color – code syntax and improved debug window and in – place object browsing.

7

System Testing

7.1 Unit Testing

7.2 Integration Testing

7.3 System Testing

Chapter 7: System Testing

Testing is the process of exercising or evaluating a system by manual or automatic means to verify that it satisfied requirements or to identify differences expected and actual results. By the other words, testing is a verification and validation process.

Verification refers to the set of activities that ensure that the software correctly implements a specific function. On the other hand, validation refers to a different set of activities that ensuring the software has been built traceable to user requirements. Software testing is a critical element of software quality assurance and represents the ultimate review of requirements specification, design and coding.

The main purpose of testing is to uncover different types of errors that exist while running the system. System testing is a critical element of software quality assurance and represents the ultimate review of specification, design and coding. A successful testing will uncover errors in the software and demonstrates that functions of a system appear to be working according to specification. However, testing cannot show the absence of defects, it can only show that software defects are present (Roger S. Pressman, 2000).

Rules that can serve well as testing objectives are:

- Testing is a process of executing a program with the intent of finding an error.

- A good test case is one that has a high probability of finding an undiscovered error.
- A successful test is one that uncovers an as yet undiscovered error.

Thus, testing is only successful when a fault is discovered or failure occurs as a results of testing procedures. The system has undergone 3 stages of testing. They are unit testing, integrating testing and system testing as shown in the Figure 7.1 below. In Figure 7.1, the arrows from the top of the boxes indicate the normal sequence of testing. The arrows returning to the previous box indicate that previous testing stages may have to be repeated because of some problems. The stages in the testing process are:

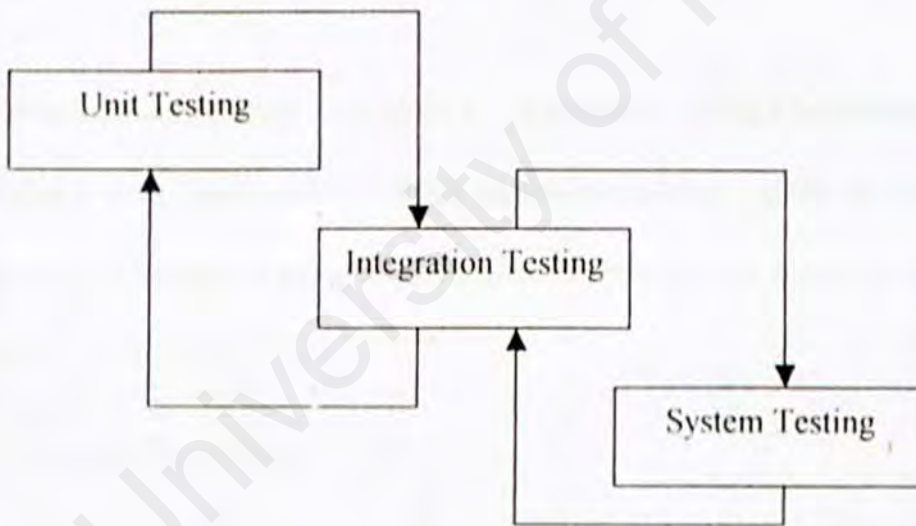


Figure 7.1: Testing Stages

7.1 Unit Testing

In this first stage of testing, each program component is tested on its own, isolated from the other components in the system. Unit testing verifies that the component functions properly with the types of input expected from studying the component's design.

For Data Mining in Electronic Commerce using Classification Technique, unit testing was done during the coding phase. The first step is to examine the program code by reading through it, trying to spot algorithm, data and syntax faults. Comparing the code with specifications and with the design to make sure that all relevant cases have been considered follows this. Finally, test cases are developed to show that the input is properly converted to the desired output.

In the development of Data Mining in E – Commerce using Classification Technique, unit testing is done concurrently with the prototyping phase. All the sub modules of Data Mining in E - Commerce using Classification Technique are tested to ensure that it is error free.

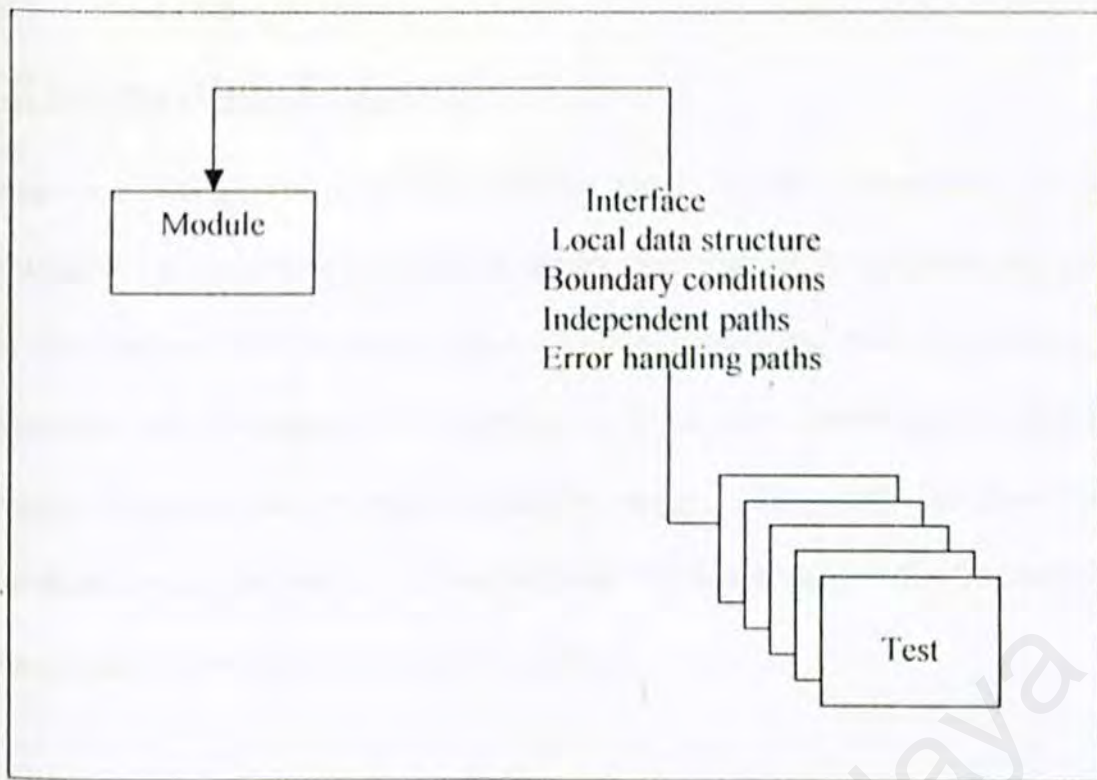


Figure 7.2: Unit Testing

The following areas are tested during unit testing for Data Mining in E – Commerce using Classification Technique: -

❖ **Interface**

Testing the interface to ensure that information flows properly into and out of the program unit.

❖ **Boundary value analysis**

Ensure that the module operates properly at boundaries established to limited or restrict processing.

❖ **Error handling paths**

Ensures that the specific module executes the recovering process should an error occurs.

❖ **All possible independent program path are executed**

Ensures that the control structures are implemented correctly.

7.2 Integration Testing

Integration testing is the process of verifying that the system components work together as described in the system and program design specification. It is a systematic technique for constructing the program structure while conducting tests to uncover errors associated with interfacing. The objective is to take unit tested modules and build a program structure that has been dictated by design. This testing will ensure that the interfaces such as the module calling sequence in Data Mining in E - Commerce using Classification Technique are arranged correctly.

In Data Mining in E – Commerce using Classification Technique, an incremental integration strategy, the bottom-up integration and regression testing approach are used. In other words, when the individual components are working correctly and meet the objectives, these components are combined into a working system. Testing the interface of 2 components explores how components interact with each other.

The incremental integration is the antithesis of the high bang approach. This system's program is constructed and tested in small segments, where errors are easier to isolate and correct; interfaces are more likely to be tested completely. Error will be corrected before processing to the next integration.

7.3 System Testing

Testing the system is very different from unit testing and integration testing. Its objective is to ensure that the system does what the users want it to do. System testing is actually a series of different tests whose primary purpose is to fully exercise the computer – based system.

System testing is designed to reveal bugs that cannot be attributed to individual component, or to the interaction among components and other objects. System tests study all the concerns issue and behaviors that can only be exposed by testing the entire integrated system or major part of it.

The Data Mining in E – Commerce using Classification Technique is tested whether it meets the specific performance testing. Data integrity testing is used to verify that the data is stored in a manner where it is not compromised under updating, restoration or retrieval processing in Data Mining in E – Commerce using Classification Technique.

System Evaluation and Conclusion

- 8.1 Problem and Solutions*
- 8.2 System Strengths*
- 8.3 System Limitations*
- 8.4 Future Enhancement*
- 8.5 Conclusion*

Chapter 8: System Evaluation & Conclusion

8.1 Problem and Solutions

8.1.1 Problem and Solution During System Studies and Analysis

8.1.1.1 Wide Area of Studies

In order to successfully develop and implement Data Mining in E – Commerce using Classification technique to be done. Furthermore, various technologies and tools had to be explored in order to choose the right tools. The Internet was a great help in obtain necessary information. Besides, knowledge was also obtained from reading of printed materials.

8.1.1.2 Determining the Project Scope

During to the time frame given, it was impossible to incorporate too many features into system. Availability of tools was also considered in determining the project scope.

8.1.2 Problems and Solution During System Implementation and Testing

As there is no prior knowledge in JAVA programming language, a lot of studies need to be done to familiarize with the concept of JAVA programming. Programming language and various development tools need to be learnt within a short time span. Choosing JAVA as the programming language was a wise decision due to this short learning

curve. Discussion with course – mates, seeking advice from Internet and self-studies also helped resolved the problems faced.

8.2 System Strengths

The system strengths that available in Data Mining in E – Commerce using Classification Technique includes the followings:

➤ Security Feature

Security is one of the important aspects of Data Mining in E – Commerce using Classification Technique. The system's security features also ensure that users information is protected from unauthorized access. This is done through the implementation of the login procedure before a user can gain access into administration section.

➤ User Friendly Interface

Data Mining in E – Commerce using Classification Technique has a very user – friendly and consistent environment. Users who are experienced in system application can easier use this system. Effective use of selection control eliminates typing need when capturing data from users. Carefully planned system make sure that users are able to navigate smoothly through system by simple point and click.

➤ Maintainability of data

Data Mining in E- Commerce using Classification Technique provides functions to enable easy maintainability where administrator can manage the database event like adding, editing and deleting.

➤ Detail user manual and navigation guides

In order to guide the users, a detailed and complete user manual and navigation guides for Data Mining in E – Commerce using Classification Technique has

been created. This will help them navigate through the system as well as understand each functions of this system.

8.3 System Limitations

➤ No Printing Capability

There is no printing facility provided in the application. Administrator and user cannot generate mining result for viewing. A more powerful printing feature should be integrated into the application.

➤ Ineffective User's Guide Linking

Due to be pressed for time, the user's guide function cannot link properly. Users have to click on 'user manual' folder to see the user's guide which the information are type by using notepad.

8.4 Future Enhancement

➤ Provide Printing Capability

Currently Data Mining in E – Commerce using Classification Technique does not support printing of information for viewing. A printing function can be incorporated to allow administrator and user to print the record retrieved from the database. This will help the user to print relevant information rather than copying them into disks.

➤ *Extra Functions*

Other interesting functions could be added to enhance its features. This will attract more users to the system and distinguish itself from other similar systems.

➤ *Pattern Implemented*

Currently analysis pattern is in text field that does not interesting. For the future enhancement, more analysis pattern can be implemented. For an example, implement result in table, figure or chart.

➤ *Effectiveness User's Guide Linking*

For the future enhancement, user's guide should be link properly.

8.5 Conclusion

The project has achieved its objectives to develop Data Mining in E - Commerce using Classification Technique which not only provides information for user but allows greater user interactivity and personalization elements. In the process, invaluable insight was gained into the complexities and intricacies of JAVA programming. Knowledge gained throughout the life cycle of project development, from the planning of the project, studies on the subject and technologies, setting up of software, programming, to implementing the system proves to be a valuable experience. At the same time, theories and knowledge gained throughout the course of information technologies studies were put into practice. This experience will definitely prove useful in future software development projects.

This is still much rooms for improvement in Data Mining in E – Commerce using Classification Technique. The successful development of Data Mining in E – Commerce using Classification Technique is the first step forwards the future development of similar systems. It is hoped that Data Mining in E – Commerce using Classification Technique can provide a foundation and basis for the concept of infotainment and its implementation using application information.

Bibliography

• Reference Book

1. Dr. Harvey M. Deitel and Paul J. Deitel. (1999). Java How To Program 3rd Ed. Prentice – Hall.
2. Dr. Harvey M. Deitel and Paul J. Deitel (1999). Visual Basic 6.0: How To Program 1stEd. Prentice – Hall.
3. Jiawei Han and Micheline Kamber (2001). Data Mining: Concepts and Techniques. London: Morgan Kaufmann
4. Kendall K.E & Kendall J.E. (1999). System Analysis and Design. 4th Ed. New Jersey: Prentice – Hall.
5. Michael J.A.Berry and Gordon S. Linoff (2000). Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: John Wiley & Sons
6. Pressman, R.S. (1992). Software Engineering: A Practitioner's Approach New York: McGraw – Hill.
7. Shari Lawrence Pfleeger. (2001). Software Engineering: Theory And Practice United States of America: Prentice – Hill.
8. Sommerville I. (1995). Software Engineering. 5th Ed. USA: Addison – Wesley
9. Kenneth E. Kendall & Julie E. Kendall (1999). System Analysis and Design, Prentice – Hall
10. P. Sellapan (2000). Systems Analysis & Design, Sejana Publishing
11. Roger S. Pressman (2000), Software Engineering: A Practitioner's Approach, McGraw-Hill

• World Wide Web

1. <http://www.microsoft.com>
2. <http://www.dbminer.com>
3. <http://www.twocrows.com/about-dm.htm>
4. <http://www.dbmsmag.com/9608d53.html>
5. <http://www.cs.bris.ac.uk/Research/MachineLearning/IBC/ilp99/node2.html>
6. <http://www.d.umn.edu/~tpederse/Pubs/naacl00.pdf>
7. <http://www.ics.uci.edu/~pazzani/Slides/BSEJ/sld010.htm>
8. <http://www.textpad.com/about/index.html>
9. <http://www-3.ibm.com/software/data/iminer/fordata/about.html>
10. <http://www.spssscience.com/clementine/index.cfm>
11. <http://georges.montefiore.ulg.ac.be/~lwh/bioinfo/notes.pdf>
12. <http://www.megaputer.com>
13. http://www.albionresearch.com/data_mining/why.htm
14. <http://www.csis.gvsu.edu/GeneralInfo/Oracle/datamine>
15. <http://www.resample.com/xlminer/help/Index.htm>
16. <http://java.sun.com/>
17. <http://www.javaworld.com/>
18. <http://www.microsoft.com/data/odbc/default.htm>
19. <http://www.cs.waikato.ac.nz/~ml/weka/>
20. www.clementine.com